DOCUMENT RESUME

ED 168 556 IR 007 090

AUTHOR	Simon, Charles W.
TITLE	Applications of Advanced Experimental Methodologies to AWAVS Training Research. Final Report, May 1977-July 1978.
INSTITUTION	Canyon Research Group, Inc., Westlake Village, Calif.
SPONS AGENCY	Naval Training Equipment Center, Orlando, Fla.
REPORT NO	NAVTRAEQUIPCEN-77-C-0065-1
PUB DATE	Jan 79
CONTRACT	N-61339-77-C-0065
NOTE	84p.; Marginally legible due to light print
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	Comparative Analysis; Cost Effectiveness; *Factor Analysis; *Flight Training; Glossaries; Program Descriptions; *Research Design; *Research Methodology; *Simulators; Transfer of Training

ABSTRACT

A major part of the Naval Training Equipment Center's Aviation Wide Angle Visual System (AWAVS) program involves behavioral research to provide a basis for establishing design criteria for flight trainers. As part of the task of defining the purpose and approach of this program, the applications of advanced experimental methods are explained and illustrated. The general AWAVS research philosophy is discussed, particularly in regard to the relative effectiveness of single versus multifactor experiments. Performance studies to be done in the AWAVS simulator are described, and a fictitious numerical example of how economical multifactor designs would be used is presented. "Quasi-transfer" experiments in the simulator are proposed to study the relationship between transfer and simulator fidelity as a composite concept rather than an entity. Some untried methods of performing multifactor transfer of training experiments more economically are suggested. A glossary of terms is included. (Author/CHV)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED168556

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

BEST COPY AVAILABLE

APPLICATIONS OF ADVANCED EXPERIMENTAL METHODOLGIES TO AWAVS TRAINING RESEARCH

Charles W. Simon

FINAL REPORT May 1977- July 1978

January 1979

Naval Training Equipment Center
Orlando, Florida 32813

Technical Report: NAVTRAEQUIPCEN 77-C-0065-1

IR007090

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NAVTRAEQUIPCEN 77-C-0065-1	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) APPLICATIONS OF ADVANCED EXPERIMENTAL METHODOLOGIES TO AWAVS TRAINING RESEARCH		5. TYPE OF REPORT & PERIOD COVERED Final Technical Report 17 May 1977 - 16 July 1978 6. PERFORMING ORG. REPORT NUMBER CWS-0178
7. AUTHOR(s) Charles W. Simon		8. CONTRACT OR GRANT NUMBER(s) N61339-77-C-0065
9. PERFORMING ORGANIZATION NAME AND ADDRESS Canyon Research Group, Inc. 741 Lakefield Road, Suite B Westlake Village, CA 91361		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NAVTRAEQUIPCEN PROJECT NO. 4781-4P3A
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Training Equipment Center Orlando, Florida 32813		12. REPORT DATE January 1979 13. NUMBER OF PAGES 80
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Aviation wide angle visual system (AWAVS)
Advanced experimental methodologies
Transfer of training research, plans
Pilot-training simulator design research, plans

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

As part of the task of defining the purpose and approach of the AWAVS human performance research program, the application of advanced experimental methods are explained and illustrated. The general AWAVS research philosophy is discussed, particularly in regard to the relative effectiveness of single versus multifactor experiments.

(continued)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Performance studies to be done in the AWAVS simulator are described and a fictitious numerical example of how economical multifactor designs would be used is presented. "Quasi-transfer" experiments in the simulator are proposed to study the relationship between transfer and simulator fidelity as a composite concept rather than an entity. Some untried methods of performing multifactor transfer of training experiments more economically are suggested.

PREFACE

A major portion of the Naval Training Equipment Center's Aviation Wide Angle Visual System (AWAVS) program involves behavioral research to provide a basis for establishing design criteria for flight trainers. Because a large number of variables will be investigated, considerable attention has been given to the methodologies appropriate for handling a problem of this complexity. Dr. Charles W. Simon has, since 1970, been studying ways in which the quality and usefulness of behavioral research can be improved through techniques that greatly increase the amount of information obtainable from a given amount of data. This contractor report summarizes his views to date concerning the application of these "advanced experimental methodologies" to the AWAVS program.

Many of Dr. Simon's technical reports, listed in the References, have not been widely distributed (although most may be obtained through the Defense Documentation Center or National Technical Information Service). Therefore, it is hoped that this report will be of benefit not only to those interested in the AWAVS program but also to those who have not yet been exposed to his work. Although not expressly intended as a primer for those unfamiliar with the research paradigm Dr. Simon advocates, portions of this report should be helpful to the new reader. In particular, Section II discusses the advantages of the multifactor approach to research, and Section V provides an illustrative example. A Glossary has also been provided.

The assistance of Dr. Daniel P. Westra is gratefully acknowledged for his critical review of this report and for his helpful suggestions.

Stanley C. Collyer
Scientific Officer

TABLE OF CONTENTS

Section		Page
I	INTRODUCTION	7
II	A MULTIFACTOR PHILOSOPHY FOR AWAVS EXPERIMENTS	9
	THE REDUCTIONISTS	9
	SINGLE VERSUS MULTIFACTOR APPROACH	10
	Definitions	11
	An Illustrative Problem	11
	SINGLE FACTOR DESIGN	12
	MULTIFACTOR DESIGN	12
	Mean Performance Measures	15
	Interaction Effects	18
	Factorial Effects in the Presence of Interactions	19
	Summary of Costs and Benefits	20
	Other Considerations	22
	SOURCES OF ECONOMY IN MULTIFACTOR EXPERIMENTS	25
III	AWAVS PERFORMANCE SIMULATOR EXPERIMENTS	28
	GROSS-EFFECT STUDY	28
	INITIAL OVERALL SCREENING EXPERIMENT	28
	VISUAL SYSTEM SCREENING EXPERIMENT	29
IV	REFINING ECONOMICAL MULTIFACTOR DESIGNS	31
	FROM RESOLUTION IV TO V DESIGNS ECONOMICALLY	31
	SCREENING DESIGNS WITH SOME FACTORS AT MORE THAN TWO LEVELS	33

TABLE OF CONTENTS (Continued)

Section		Page
V	APPLYING ECONOMICAL MULTIFACTOR DESIGNS TO AWAVS PERFORMANCE EXPERIMENT -- AN EXAMPLE . . .	35
	EXPERIMENTAL OBJECTIVES	35
	EXPERIMENTAL FACTORS	35
	EXPERIMENTAL SUBJECTS	35
	EXPERIMENTAL PLAN AND PROCEDURE	36
	Analysis of the First Set of Data . . .	40
	Obtaining a Response Surface	47
	VERIFICATION AND FIDUCIAL LIMITS	50
VI	QUASI-TRANSFER EXPERIMENTS	51
	FIDELITY	52
	Dimensionalizing the Situation	52
	Dimensionalizing Fidelity	54
	Experiments	54
	Novel Transfer of Training Designs . .	56
	AWAVS AS A CRITERION DEVICE	59
VII	ECONOMICAL DATA COLLECTION PLANS FOR TRANSFER OF TRAINING STUDIES FOR THE AWAVS PROGRAM	60
	PERFORMANCE TO TRANSFER APPROACH (I)	61
	Selected Configurations (Plan I-A). . .	61
	Performance-Transfer Prediction Approach (Plan I-B)	65
	LIMITED DIRECT TRANSFER APPROACH (II). . . .	66
	Complete Transfer Surface (Plan II-A) .	67
	Search-For-Optimum Transfer (II-B). . .	68
	Reduced-Flight Predicted Transfer (II-C)	69
VIII	SOME UNFINISHED BUSINESS - MEASUREMENTS AND CRITERIA	71
	REFERENCES	73
	GLOSSARY	75

LIST OF ILLUSTRATIONS

Figure		Page
1	Location of Data Collection Points in a Five Dimensional Space	17
2	Half Normal Plot of Experimental Data in Table 5.	44
3	Theoretical Relationships Between Transfer, Simulation Fidelity and Costs.	53

LIST OF TABLES

Table		Page
1	Experimental Design Typical of One-Factor Experiment	13
2	Typical Multifactor Experimental Design (for Twelve Factors)	14
3	Comparing Single Factor and Multifactor Approaches	21
4	Data for 2^{12-7}_{IV} Trend Resistant Screening Designs, N = 32	37
5	Analysis of Fictitious AMAVS Data from Design in Table 4	41
6	Three Factor Interactions in the Critical Strings in Table 5	45
7	Examples of Dimensions of Fidelity in Visual and Motion Simulation Systems	55
8	Approaches to Economical Transfer of Training Research	62

SECTION I

INTRODUCTION

The Naval Training Equipment Center is building a sophisticated pilot training simulator which focuses on advancing the state-of-the-art of the visual system. The aviation wide angle visual system (AWAVS) along with a six-degrees of-freedom motion system combine to provide a highly versatile simulator on which complex behavioral research can be performed. Initially, the primary purpose of such research will be to examine and optimize the simulator parameters for pilot training in specific carrier-landing tasks. The large number of parameters that must be investigated requires the use of advanced experimental methodologies for studying many factors economically. A discussion of philosophy, strategy, and techniques that might be employed on this program represents the basis for this report.

Two types of investigations have been proposed for research on the AWAVS simulator. These will be referred to as "performance" experiments and "transfer" experiments. A "performance" experiment is one that measures operator/system performance under one set of conditions, presumably uninfluenced by any other conditions. Measuring pilot performance in an aircraft under different instrument conditions or in a simulator with different configurations could be an example of this type of experiment. A "transfer" study is one in which the interest is in the residual effect that practice on one set of conditions has on the performance of a second set of conditions which follows it in time. In this report, two classes of "transfer" experiments are defined. A "real transfer" (referred to as "transfer") experiment for the AWAVS task is one in which the training occurs in the simulator while the test of residual transfer occurs in flight in an aircraft. A "quasi-transfer" experiment for the AWAVS task is one in which both pilot training and transfer testing (representing flight) occurs in the simulator.

Previous work on this program had emphasized the planning of the performance experiments, the type to be performed first when the AWAVS simulator is operational. In this report, more emphasis is placed on developing new and economical ways in which transfer experiments might be performed, to enhance the pragmatic value of results from such experiments.

This report is not a review of the literature. Its purpose is to increase the understanding of those less familiar with "advanced experimental methodologies" as they might be applied to the AWAVS program. It will also briefly summarize the conceptualization of new, economical approaches that might be employed to aid in the understanding and measure of transfer of training for the carrier-landing task. Detailed explanations will be avoided here. For a background in "advanced experimental methodologies," the reader may wish to refer to reports prepared by Simon (1970 through 1977).

The following topics will be treated in this report:

a. A multifactor philosophy for AWARS experiments

b. AWARS performance simulator experiments

c. Refining economical multifactor designs

d. Applying economical multifactor designs to AWARS performance experiments -- an example

e. Quasi-transfer experiments

f. Economical data collection plans for transfer of training in the AWARS simulator

g. Some unfinished business - measurement and criteria.

SECTION II

A MULTIFACTOR PHILOSOPHY FOR AWAVS EXPERIMENTS

The philosophy of AWAVS experiments differs from that employed in other training simulation design and transfer of training experiments. For the AWAVS studies, a "holistic" philosophy has been accepted as categorically imperative. This philosophy espouses the need to include in experiments, during the factor identification and function development phases of a research program, as many factors as possible that are believed critical to the particular operational task under investigation. The more one is able to achieve this goal, the less likely the data will be biased, the more accurately laboratory data will predict the operational situation, and the more readily a quantitative, modular data base for application to future problems can be built (Simon, 1977b). Until attention was focused on the various techniques and paradigms for conducting systematically controlled large scale multifactor experiments economically, the size of the effort was a limiting feature to this holistic philosophy. The general approach that is proposed for the AWAVS experiments makes this no longer a critical consideration.

The novelty of the proposed approach lies primarily in the economical patterns -- both spatial and temporal -- employed in selecting the points forming the simulation space that corresponds to those in an operational situation. Advanced techniques are also used to keep the information of primary interest unconfounded with effects from irrelevant sources and to do so without disrupting the economy of the effort. The quantity and quality of information from this multifactor approach in almost every respect exceeds that obtained by other techniques used by psychologists employing the same amount of resources.

THE REDUCTIONISTS

That some do not fully adhere to this philosophy in the conduct of behavioral research is evidenced by a report recently prepared by a working group of the Vision Committee of the NAS-NRC. For pilot training research at Williams Air Force Base, this group recommended ways "to increase the effectiveness of experiments" on visual cues in flight simulators. Their number one recommendation was:

> Simplify the experimental design whenever possible. Attempt to identify the major parameters with exploratory studies and then examine these parameters one at a time rather than using a multifactor design. (NAS-NRC, 1976, p.9)

Earlier in that report they had listed a number of parameters that should be investigated in an initial evaluation of the realize issue, and concluded that "the interactions between major parameters should also be studied, but only at a later date after the effects upon task-specific training have been determined by varying one parameter at a time."

The report was in draft form when it was seen and efforts to find a final copy have been unsuccessful. However, the issue here is not with the report per se; it is mentioned only to illustrate the fact that the one-factor-at-a-time approach to behavioral research still has its adherents, even among prestigious groups with considerable influence on the nature of major research programs. Consequently, the relative merits of single and multifactor approaches must be examined.

SINGLE VERSUS MULTIFACTOR APPROACH

In the remainder of this section, a comparison of two approaches will be made as they are applied to the task of identifying critical factors and measuring their effects, and deriving an equation to predict performance under operational conditions. A candidate list of twelve factors will be used to illustrate how the information/cost ratio is affected by experiments employing each approach. "Cost" here refers to data collection cost.

In this discussion, the following claims will be supported:

a. Given the same time and resources, the multifactor approach will always provide quantitatively and qualitatively better* information than a single (or few) factor approach will.

b. There is certain information that a single factor approach can never provide, but which is available when a multifactor approach is used.

---

* Information is judged "better" when it has more of the following qualities:

- economy in data collection
- precise estimates within experiment
- accuracy when producing from laboratory data to operational situation
- ability to generalize to numerous situations
- ability to use the data to construct a modular data base for future reference
- ease of smoothing faulty data
- reduced ambiguity in interpretation

Definitions

The number of factors in an experiment can range from 1 to N. References in this report to single or multifactor experiments therefore are associated with opposite ends of that continuum. Traditionally, most behavioral research in equipment design has included only two or three factors in a single experiment, and prior to the use of "economical multifactor designs," an experiment with more than five factors was a rarity (Simon, 1976b). In this section reference to a single factor or one factor experiment implies a class of experiments recommended by reductionists who believe that good behavioral information can be obtained by studying one factor at a time. However, most comments made here regarding this class of experiments will sometimes apply, to a lesser degree, to experiments involving two, three, four, and even five factors, when a great many more factors are in fact critical for the particular task under operational conditions. Reference to a "multifactor" experiment implies that it entails an effort to include most of the candidate factors believed to influence the behavior found in the particular investigation. Merely including more than one factor in an experiment would not meet the requirement of a multifactor experiment as the term is used here.

In practice, there are usually only a relatively few highly critical factors affecting performance on a particular task. However, to include most of the critical factors in an experiment, it is usually necessary to start with a much larger number of candidate factors. It is assumed that for most behavioral problems, persons working in the field can identify candidate factors that have the potential for influencing the class of behavior under investigation, but that for any particular task, only an empirical effort can determine how much effect each factor has, and therefore which ones are critical. While we may never achieve a one-hundred percent inclusion of critical variables in a controlled experiment, we can at least increase considerably the number over that which tends to be typical in experiments today.

An Illustrative Problem

Let us look at one of the experiments that might be done on the AMAVS program and compare what would happen if a single factor or a multifactor approach were used. Let us assume there are twelve simulator factors plus one pilot and one task difficulty (environment) factor, all at two levels each. For the time being, we shall not include the last two, since that would only complicate the discussion without altering the conclusions. The purpose of the experiment is to find out which of the twelve factors will be critical in the design of a pilot training simulator (using simulator performance as the criterion) and what performance levels each of the two conditions (levels) of each factor yields.

SINGLE FACTOR DESIGN

The typical single factor approach might follow this design. Select one factor -- Factor A -- and test eight pilots on the one condition of Factor A and eight other pilots on the other condition of Factor A. Pilots would be assigned to each group at random. The remaining eleven simulator factors would be held constant as, presumably, would all irrelevant sources of variance. This design is illustrated in Table 1*.

When the data has been collected the mean performance for each of the two conditions (levels) of Factor A can be calculated and the effect of Factor A, i.e., the difference between these two means, can be estimated. The precision with which each effect is estimated, i.e., the standard error of the mean difference (σ_{md}), can also be calculated. The equation for this illustrative problem is:

$$\sigma_{md} = \sqrt{\frac{\sigma^2_+}{N_+} + \frac{\sigma^2_-}{N_-}} = \sqrt{\frac{2\sigma^2}{8}} = .5\sigma$$

where σ^2 is the estimated error variance of the experimental unit (independent of factors), and N is the total number of observations made per experimental condition. Once the appropriate σ is established, this standard error of the mean difference can be used to set confidence limits about the empirically determined means.

MULTIFACTOR DESIGN

Using the multifactor approach, the effects of all twelve factors would be estimated in a single experiment also composed of a total of 16 observations from 16 pilots, one per observation.

---

* Slightly modified experimental designs have been used in "one factor" experiments. For example, a subject (pilot) might be tested on both experimental conditions. To compensate for carry-over effects, one-half the subjects would be presented the conditions in one order, and one-half in the opposite order. For our discussion, these variations are not critical. While only eight subjects would be required, the total number of observations remains 16, and it is the number of observations, not pilots, that will be the unit of measuring the cost effectiveness of the data collection.

TABLE 1. EXPERIMENTAL DESIGN TYPICAL OF ONE-FACTOR EXPERIMENT

	FACTOR A	
	Condition 1 (-)	Condition 2 (+)
Pilots	1	9
	2	10
	3	11
	4	12
	5	13
	6	14
	7	15
	8	16

FACTORS HELD CONSTANT	
Factor	Value*
B	(Single value for each selected at experimenter's option.)
C	
D	
E	
F	
G	
H	
I	
J	
K	
L	

*Value refers to one condition or the other, designated - or +. With quantitative values, these would correspond to low or high levels, and be a shortened notation of -1 and +1. The values at which each factor is held constant would be decided by the investigator.

The experimental design for this 2^{12-8}_{III} experiment is shown in Table 2-A. The minus and plus signs in the table represent the high or low (or first or second) level of each factor. Each row represents a different experimental condition and each column -- up to twelve -- a different factor. With this design, the main effects of <u>all</u> twelve factors can be estimated. The precision with which each one of the main effects can be estimated with this design is the same as the precision of the effect estimates in the single factor study, namely .5σ*. Thus, finding the main effects of twelve factors with the single factor approach would cost twelve times as much as with the multifactor

*In this example, the multifactor design is not replicated; therefore, there is no direct estimate of the "error" standard deviation (σ_e). Internal estimates can be made, however, from the half-normal plot as shown on page 44 (see Simon, 1977, p 97).

TABLE 2. TYPICAL MULTIFACTOR EXPERIMENTAL DESIGN (FOR TWELVE FACTORS)

Conditions	MEAN	DL CK BJ EG A	HL FK AJ EI B	IL AK FJ EH C	AL IK HJ EF D	BI CH AG DF E	GL BK CJ DE F	FL HK IJ AE G	BL GK DJ CE H	CL DK GJ BE I	GI DH CF AB J	DI GH BF AC K	AD CI BH FG L	EL BC JK HI DG AF (ACJ)	JL BD EK FI AH CG (ADJ)	KL CD EJ AI FH BG (ADK)
							(TABLE 2-A. FIRST BLOCK)									
1. EJKL	+	-	-	-	-	+	-	-	-	-	+	+	+	+	+	+
2. AFHI	+	+	-	-	-	-	+	-	+	+	-	-	-	+	+	+
3. BFGHKL	+	-	+	-	-	-	+	+	+	-	-	+	+	-	-	+
4. ABEGIJ	+	+	+	-	-	+	-	+	-	+	+	-	-	-	-	+
5. CFGIJL	+	-	-	+	-	-	-	+	-	+	+	-	+	-	+	-
6. ACEGHK	+	+	-	+	-	+	-	+	+	-	-	+	-	-	+	-
7. BCEHIL	+	-	+	+	-	+	-	-	+	+	-	-	+	+	-	-
8. ABCFJK	+	+	+	+	-	-	+	-	-	-	+	+	-	+	-	-
9. DGHIJK	+	-	-	-	+	-	-	+	+	+	+	+	-	+	-	-
10. ADEFGL	+	+	-	-	+	+	+	+	-	-	-	-	+	+	-	-
11. BDEFIK	+	-	+	-	+	+	+	-	-	+	-	+	-	-	+	-
12. ABDHJL	+	+	+	-	+	-	-	-	+	-	+	-	+	-	+	-
13. CDEFHJ	+	-	-	+	+	+	+	-	+	-	+	-	-	-	-	+
14. ACDIKL	+	+	-	+	+	-	-	-	-	+	-	+	+	-	-	+
15. BCDG	+	-	+	+	+	-	-	+	-	-	-	-	-	+	+	+
16. ABCDEFGHIJKL	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
							(TABLE 2-B. SECOND BLOCK)									
17. ABCDFGHI	-	+	+	+	+	-	+	+	+	+	-	-	-	-	-	-
18. BCDEGJKL	-	-	+	+	+	+	-	+	-	-	+	+	+	-	-	-
19. ACDEIJ	-	+	-	+	+	+	-	-	-	+	+	-	-	+	+	-
20. CDFHKL	-	-	-	+	+	-	+	-	+	-	-	+	+	+	+	-
21. ABDEHK	-	+	+	-	+	+	-	-	+	-	-	+	-	+	-	+
22. BDFIJL	-	-	+	-	+	-	+	-	-	+	+	-	+	+	-	+
23. ADFGJK	-	+	-	-	+	-	+	+	-	-	+	+	-	-	+	+
24. DEGHIL	-	-	-	-	+	+	-	+	+	+	-	-	+	-	+	+
25. ABCEFL	-	+	+	+	-	+	+	-	-	-	-	-	+	-	+	+
26. BCHIJ	-	-	+	+	-	-	-	-	+	+	+	+	-	-	+	+
27. ACGHJL	-	+	-	+	-	-	-	+	+	-	+	-	+	+	-	+
28. CEFGIK	-	-	-	+	-	+	+	+	-	+	-	+	-	+	-	+
29. ABGIKL	-	+	+	-	-	-	-	+	-	+	-	+	+	+	+	-
30. BEFGHJ	-	-	+	-	-	+	+	+	+	-	+	-	-	+	+	-
31. AEFHIJKL	-	+	-	-	-	+	+	-	+	+	+	+	+	-	-	-
32. (1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

* In the first block, main effects are aliased with two-factor interactions as shown, along with higher-order interactions. Where no main effects are shown, one of a string of three-factor interactions is shown in parentheses.

When data from the second block is added to that from the first block, main and three-factor interaction effects are isolated from the strings of two-factor interactions.

Block I and Block II are each Resolution III designs. Combined they form a Resolution IV design.

study and the precision of each estimate would be no greater. Conversely, the main effects of twelve factors can be estimated at the same data collection cost with the multifactor approach as one can estimate one factor with the single factor approach, and with the same precision.

Mean Performance Measures

There are still more important and subtler differences between the two approaches that are often overlooked. For example, the means obtained with the single factor study will be different from the means obtained in the multifactor study. This is so in spite of the fact that both were obtained by measuring performance eight times at the high level (of Factor A) and eight times at the low level (of Factor A). Unfortunately, the means obtained from the single factor study are not representative of performance throughout the experimental space. Instead, the two means are obtained by measuring only two locations out of a possible 4096 in the total experimental space (in this example). These two locations are at the edge of the twelve-dimensional hypercube, representing less than five ten-thousandths of the full factorial space.

But it is not the small proportion that is critical, per se; it is the fact that these means estimate are not independent of the factors held constant. In spite of many replications and what might appear to be a very uncomplicated experimental design, the chances of obtaining a reasonably accurate estimate of the performance on either the high or low condition of Factor A (in our example) is very poor when the single factor approach is used. This is because the answers we obtain with such a design depend on which values the investigator decides to use for the factors held constant. Because they are held constant does not mean that they have no effect on performance; they do.

If a factor that is held constant would critically affect performance were it varied, then the value at which it is held constant will make the overall task either easier or more difficult to perform. Thus, mean performance on the conditions of Factor A would increase or decrease from the average, depending on the particular values at which the constant factors are fixed by the investigator. In the single factor study, the combination of fixed values is only one out of a possible 2048 alternatives (i.e., one out of 2^{11} combinations). Since the single factor experiment tells us nothing of the effects of these factors, we have no way of knowing in which direction the bias lies nor its magnitude.

With a multifactor approach, the situation is different. The means for Factor A are more representative since they are obtained by sampling a number of conditions throughout the

experimental space. The other factors are not held constant but are varied systematically and orthogonally to one another as well as to Factor A (see Table 2-A). Each level of Factor A is measured in combination with an equal number of high and low conditions of every other factor, thereby neutralizing the effects of the other factor on the mean performance for each condition of Factor A. The same balance occurs with all other factors in the multifactor experiment.

Graphic example. The above relationships may be more easily understood if they are shown graphically. Since it is difficult to draw a twelve-dimensional space on two-dimensional paper, let us use a five-dimensional space to illustrate what has been said so far. In Figure 1, two diagrams each representing a five-dimensional space are shown. The one on the left will be used in the discussion of the single factor design and the one on the right, of the multifactor design. In the diagram on the left, at the corners of each cube, the thirty-two conditions of a five factor, two levels per factor space, are identified. The conditions would be identically named in the corresponding positions on the right. The conventional symbology for naming experimental conditions is employed, where the presence of a letter, a through e, indicates that the high (+) level of factors A through E respectively is represented. The absence of a particular letter indicates that the low (-) level of that factor is represented in the condition. Black dots have been imposed on each diagram where data is to be collected.

In the single factor experiment, two conditions at which performance under the high and low levels of Factor A are to be compared are selected arbitrarily, i.e., bc and abc*. Note that in this five factor case, any one of 16 alternatives could have been chosen, all of which run only along a horizontal edge of a cube in Figure 1. Once the two conditions are chosen, eight measurements are made at each condition. However, if Factor C has a large effect on performance, with the + condition causing the higher performance level, then the means at bc and abc would be higher than if the single factor study of Factor A had been carried out with Factor C being held constant at its lower level, e.g., conditions e and ae. This process becomes even more complex if other constant factors also had critical effects. Even after data has been collected through a series of single factor experiments on all the factors,

---

*Any pair of conditions could have been selected as long as a is absent from one and present in the other and all other letters are held constant, i.e., the same in both. For example, bce and abce, e and ae, and so forth, could also have been used.

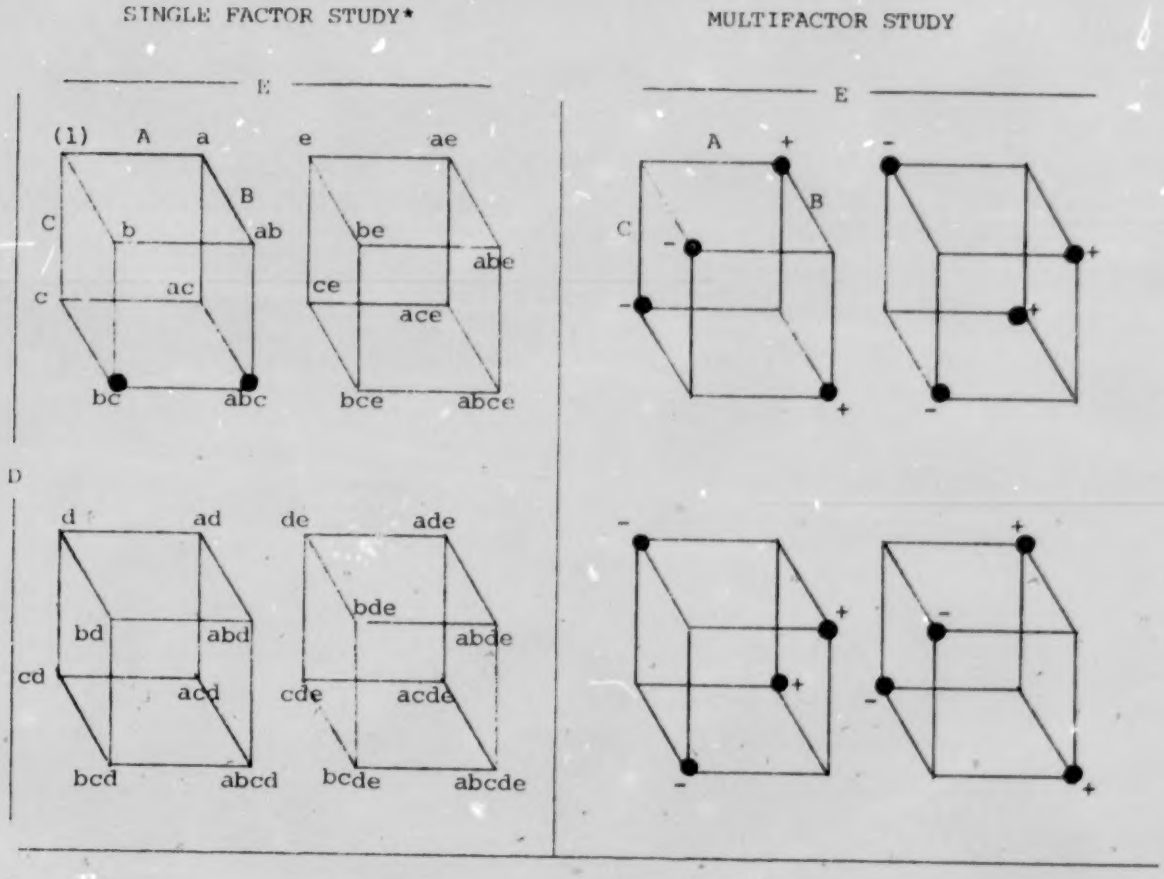

Figure 1. Location of Data Collection Points (●) In A Five Dimensional Space.

*Each point in the Single Factor study is replicated eight times. This makes a total of 16 observations. In the multifactor study, the 16 observations are distributed as shown with no replication.

there is still insufficient information to correct the mean estimates for potential biases. The fact that the individual one factor studies were performed sequentially without any method of correcting or measuring possible sequential effects that would cause irrelevant variations in performance from study to study makes any estimates of mean performance even more suspect.

In the multifactor experiment, the data points (shown in Figure 1) were selected to prevent the mean performance values from being affected by the other factors in the experiment.

The eight points marked "minus" represent the low condition of Factor A and the eight points marked "plus" represent the high condition of Factor A. Means of the measures obtained for the minuses and for the pluses would represent the averages for the two levels of Factor A. Note that half of the low level points for Factor A were measured under a high condition of every other factor and half were measured under a low condition of every other factor. Any effect that these other factors might have on task difficulty has been balanced out in the estimation of Factor A in this multifactor plan. The same would be true were the means of any other factors estimated. The main effects of each are orthogonal to one another*.

Interaction Effects

With no interactions among the factors, even the single factor approach will arrive at an appropriate estimate of the effect of Factor A. This is true even though the means of each condition, as previously illustrated, may be higher or lower than what their "true" value would be because the factors held constant are at values that make performance easier or more difficult. When there is no interaction, since both means are affected the same, the difference between them would remain constant whatever the effect of a fixed factor.

For example, in a single factor study, if eight measures were taken each on the high and the low condition of Factor A, and if Factors B and C are each held constant at their high level and Factors D and E, at the low level -- the data collection points bc and abc indicated in Figure 1 -- the following fictitious data might be obtained:

MEANS OF CONDITIONS: bc = 23, abc = 32;

EFFECT OF FACTOR A = +9

If, instead, the high and low conditions of Factor A were compared when only Factor B was held constant at the high level and Factors C, D, and E were each held constant at their low level, and if Factor C actually had a strong effect on performance,

---

*One alternate plan exists for this example. The undotted points might have been used instead of the dotted points with the same results. This happens to be a 2^{5-2} fractional factorial. Note that points in this design are always located on diagonals, in contrast to the horizontal location of points in the single factor experiment.

e.g., ∓2, then a change from the high condition to the low condition of Factor C when it is held constant, would cause the performance scores to change by -4 points, i.e :

MEANS OF CONDITIONS: b = 19, ab = 28

EFFECT OF FACTOR A = +9

The means dropped, but in the absence of interaction effects, the effect of Factor A is unchanged.

Factorial Effects in the Presence of Interactions

But if there are interactions then we cannot trust the estimates of the effects in either the single or multifactor approach. However, there is a difference, as we shall show; the multifactor approach can handle this problem while the single factor approach cannot. To illustrate this, let Factor C interact with Factor A such that when the high levels of both occur in the same experimental condition, performance is improved far beyond what would be expected from a linear combination of the effects of each factor alone. Arbitrarily let us say it adds nine points to the mean of that condition.

In a single factor experiment, we might get these results if Factor A were studied with Factors B and C held constant each at their high levels and Factors D and E at their low levels, e.g., the marked conditions in Figure 1:

MEANS OF CONDITIONS: bc = 23, abc = 32

EFFECT OF FACTOR A = +9

But if the investigator had by chance chosen to hold Factor B at the high level and Factors C, D, and E at the low levels, these results might have been observed:

MEANS OF CONDITIONS: b = 23, ab = 23

EFFECT OF FACTOR A = 0

The interaction effect, when the high conditions of Factors A and C occurred in the same experimental condition together, (as in condition abc), made Factor A appear to be a critical effect. But had the investigator used conditions b and ab with Factor C held constant at its low level, the results would have led to the conclusion that Factor A was not an important factor in equipment design. Thus he might decide to omit it later in an interaction study. In the study with 12 factors there are 2048 pairs of points to choose from, of which -- if Factor C were the only consideration -- one set of 1024 would have led to the conclusion that Factor A was trivial and the other 1024 to the conclusion that Factor A was critical. A 50-50 chance, some

may feel, are good odds. Competent experimenters do not rely on chance. The only way to estimate two factor interactions is to include both factors in the experiment.

In a multifactor experiment, the situation is somewhat different. It is true that with the design shown in Table 2-A, wherein 12 main effects can be estimated from only 16 experimental conditions, all main effects are completely confounded with some two factor and higher order interactions and suffer the same ambiguities of interpretation as in the single factor experiment. However, by adding only 16 additional experimental conditions to the design (see Table 2-B), we can isolate the twelve main effects from all two factor interaction effects. Although we have doubled our original allotment of only sixteen observations to achieve this, we have still used only one-sixth the effort required to study all twelve factors with a single factor approach. What is more, in the single factor experiment, the main and two factor interaction effects would still remain confounded.

Summary of Costs and Benefits

Let us summarize what has been found out regarding the two approaches up to this point. What are the costs and benefits of using each approach -- single and multifactor -- to determine the relative importance of twelve factors? To achieve this, it is necessary, as a minimum, to determine the effect of each factor, isolated from two factor interaction effects, but with critical interactions identified. Table 3 summarizes costs and achievements described up to this point.

Therefore, with 32 observations in the multifactor experiments we can study the effects of 12 factors with even greater precision than we would have obtained with 192 observations required in the series of single factor experiments. Furthermore, for each new factor to be studied using the single factor approach, another increment of observations are required, in our example, an additional 16. Using this multifactor design, this is not the case. To isolate the main from all two factor interactions, we can study up to 16 factors with 32 observations, no more than were required to study 12. It would require only 64 observations -- still fewer than the number required to study 12 factors by the single factor approach -- to study up to 30 factors with main effects isolated from two factor interactions.

In both approaches, main effects can be biased by three factor interactions. More data must be collected to isolate these, if necessary, when multifactor designs are employed. No recourse is possible with the single factor study, which can only start over again -- without clues -- and do a multifactor study to discover and isolate interactions. The

TABLE 3. COMPARING SINGLE FACTOR AND MULTIFACTOR APPROACHES

	Single Factor	Multifactor
Number of main effects isolated from one another	12	12
Mean estimates of experimental conditions are unaffected by level of other factors if there are no interactions	No	Yes
Precision with which each effect is estimated*	.50σ	.35σ
Number of main effects isolated from two factor interactions	0	12
Estimates of main effects are affected if two factor interactions are present	Yes	No
Detect the presence of two factor interactions or clues as to where two factor interactions might exist	No	Yes
Main effects confounded with three factor interactions	Yes	Yes
Total number of observations used to achieve this	192	32
Planned capacity to further expand experimental space by augmenting existing data	No	Yes

*Each effect in each single factor experiment was estimated with 16 observations. Each effect in the augmented multifactor experiment was estimated with 32 observations.

multifactor approach builds on the original data. Since the original single factor studies can be biased for the reasons cited earlier, they cannot even supply data that might reduce the size of subsequent multifactor studies, i.e., the number of factors needed to be included in the multifactor studies. It would be risky to eliminate a factor based on single factor study information.

Without replicating the experimental designs, an act that would reduce its economical quality, the multifactor approach has no direct method of measuring error variance, and therefore cannot make a traditional test of statistical significance of the differences. With the single factor approach, the within-cell subject variability is conveniently labelled "error" variance and the mechanics of a test of statistical significance can be followed. This does not reduce the effectiveness of the multifactor approach, however, for several reasons. For one thing, the test of statistical significance as it would be applied here is of limited value in the interpretation of the data (Simon, 1971; 1977b; NTSC, 1976). For another, there are other equally effective methods of examining whether observed effects are the result of "chance" or not when the multifactor approach is used. One of these, i.e., half-normal plots, is illustrated in Section V. Economical partial replication techniques are also available.

Other Considerations

There are less tangible but equally important reasons for considering only a multifactor approach in equipment design research. When a multifactor approach is used -- and we have shown that it is much more economical -- the information obtained will be more generalizable, will explain more, will be easier to interpret, and will enable more accurate predictions to be made from the laboratory data to operation situations.

Generalizability. Multifactor approaches are more generalizable by the very fact that they investigate more conditions of more factors. Given the results from one of these experiments, an investigator may consider a wide range of alternative simulator parameters; to be truly generalizable, the experiment must also include contextual factors. For example, pilot training simulator studies have sometimes been critised because they used pilots with one kind of experience to obtain data that was applied to situations in which the pilots have different kinds of experiences; or in studies done under simulated conditions for low performance aircraft when the results were applied to situations in which high performance aircraft would be involved. While this is more the fault of the user than of the experimenter, still it raises the question of whether or not non-representative experiments can be justified at all? Simple experiments lack generalizability; multifactor experiments can

achieve more generalizability by including not only the simulator parameters, but others associated with pilot, task, and environment characteristics. If introduced at the beginning of the research program, during factor identification stage, they can be studied far more economically and enable more generalizable results to be obtained.

Component contributions. The multifactor approach can also provide better information than the single factor approach in situations where complex devices are being studied, as in the case of a pilot training simulator. While one may think of the visual or motion system as unitary components, results may be totally misleading when components as complex as these are treated as units. Each is made up of sub-components which have their own individual effects on performance or on transfer of training. A motion/no motion study is a case in point. Motion in a simulator can serve two relatively diverse purposes: 1) it can provide the pilot kinaesthetic cues he may use to better control his aircraft; or 2) it can simulate environmental disturbances that can negatively affect the ease with which the aircraft can be controlled. Combining these two purposes may not have the same effect on training. A study in which these effects are not examined separately, as two independent factors, might lead to the conclusion that there is no overall difference between a motion or a no motion system, if the effects of these two components were in fact in opposition and cancelled one another. A similar illustration might be used in regard to the study of a motion system in which several motion cues are used, e.g., simulator movement and G-seats. Unless they are studied separately (and the multifactor approach is the cheaper way of doing this), their effects might cancel one another. Similarly, a comparison of two simulator configurations to see which is the better might suffer from this same problem, e.g., the existence of a superb visual system in one configuration and a superb motion system in the other configuration might lead to a stand-off, showing both to be similar in effectiveness and never revealing which combination might have produced the super-simulator so long sought after.

Interpretation. When only two data points are investigated, the investigator has no way of evaluating the correctness of the results through rational processes. When a great many data points are collected in the systematic manner of the multifactor design, the investigator has built-in checks in the form of data patterns. Erratic behavior is more likely to be spotted, giving the investigator the opportunity of checking whether it is an outlier or a bona fide interaction.

A multifactor approach also puts the interpretation of experimental results in perspective. When a single factor is studied alone, it is more difficult to judge its relative importance to the system. Importance is more clearly evident

when the proportion of variance a factor accounts for is known relative to that accounted for by all of the other primary factors affecting a particular task. When allocations of time and money require that improvements in equipment design be considered on a priority basis, knowledge of one factor's effect on system performance in context with all others is an important interpretative feature provided best by the multifactor approach.

Prediction. The experimental designs traditionally employed by experimental psychologists have been more concerned with precision of results rather than accuracy. Precision refers to the reputability of a measure, whether it is biased (inaccurate) or not. The single factor approach, as has been shown, maximizes bias and obtains a satisfactory level of precision only at considerable cost. The multifactor approach (with a holistic philosophy) emphasizes the reduction of bias and at the same time, because of its inherent features, tends to maintain precision quite economically. The relative merits of the single versus multifactor approach were discussed by the eminent statistician, Frank Yates (1970, p 51, more than forty years ago. At that time, he made the following comment: ". . . the experimenter who confines himself to experiments on single factors, making a guess at the final levels of the other factors, is merely emulating the tactics of an ostrich."

Because we can include in our experiments most of the factors critical under operational conditions, as well as those affecting the pilot, task, and environment, the multifactor approach increases the accuracy of our predictions. When the single factor approach is used, each critical factor omitted (hence constant) from an equation can break a prediction if it does not match that found operationally; each one that is allowed to vary in the experiment results in variable prediction error. Even at the end of the experimental program when only a few configurations might be examined for purposes of verification, detailed comparison, or for establishing fiducial limits on the performance, the multifactor approach has already provided an overall framework into which the data from the limited experiment can be anchored.

The use of a sequential block technique for data collection in the multifactor approach can help optimize prediction. If the investigator has reason to suspect that the order of his predictive model is inadequate, i.e., would fail to fit reality, he may collect additional data that would be combined with the original data so as to enable quadratic or higher-order surfaces to be estimated if necessary.

SOURCES OF ECONOMY IN MULTIFACTOR EXPERIMENTS

Historically, accepting the need to perform holistic (multifactor) experiments has proven to be easier said than done. In 1954, for example, Williams and Adelson, wishing to examine the effects of 34 factors they believed important in the design of a pilot training simulator, were stymied by the fact that a factorial design for 34 factors at five levels each would require 5.8×10^{23} combinations. Even studying each factor, one at a time, at five levels with all other factors held constant, would have required 3400 observations. To reduce the effort they considered studying only the important factors, but recommended that no study be done at that time since the original 34 had been selected because they were the important ones. The same questions regarding pilot training simulators and a method of doing a comprehensive experiment continue to exist during the intervening 25 years. Simon (1970a, b, 1971, 1973, 1974, 1977a, b) proposed a more economical approach with which to accomplish this task. A few of the more important principles for achieving this economy are cited here briefly.

First of all, it is not necessary to collect data with which to isolate higher-order interactions. In the example cited above, it is a certainty that no 34-factor interaction would be of any practical importance. For that matter, no ten-, or six-, and probably no four factor interaction will have a practical effect on performance. Even three-factor interactions seldom have large effects, particularly if quantitative, continuous factors are involved (Simon, 1976b). To illustrate the savings this observation can achieve, let us consider a 15 factor study. A complete factorial for 15 factors would require 32,768 combinations if each factor were studied at two levels, or 14,348,907 combinations if each were studied at three levels. However, if the response surface for 15 factors could be represented by a first-degree equation, only 16 properly selected conditions would be required. If it could be represented by a second-degree equation, then only 136 conditions would be required. If the surface could be represented by a third-degree equation, then 816 conditions would be required. While the latter number of conditions is still large, it is only a .000057th fraction of the complete 3^{15} factorial.

To be economical, however, an experiment would never be started with the intention of measuring 816 conditions, even if we thought that a third-degree surface need be represented. We would begin by collecting only enough data to approximate a first-degree surface. Then a little additional data would be collected in order to test whether this first-degree approximation adequately fits the response surface. If it does, the study can stop, thereby achieving considerable economy. If not, additional data would be collected to approximate a second-degree surface and a second test would be made. If the fit is adequate the study would stop at this point; if not, it would continue. This iterative

process serves two purposes: one, it keeps the cost of the experiment as low as possible; two, it provides the assurance that the response surface will be adequately represented. Theoretically, the procedure could continue up through fourth and fifth-degree surfaces, although this is highly unlikely with psychological data. Interactions at that level would more probably indicate that the data were carelessly collected or that the experimenter had failed to scale his data properly.

Proper scaling is another way to achieve economy in multifactor experiments. Certain classes of interactions and curvature can be eliminated by selecting the appropriate scale. If care is taken before the data are collected to select the correct scales, the necessity of approximating a third- or even a second-degree surface is diminished and less data need be collected. Certain interactions cannot be avoided by scaling, but in the behavioral sciences these occur infrequently.

Still further economy can be achievedif we separate the critical factor identification process from the function derivation process. Why should we collect the data required to develop a third- or second-degree function for 15 factors if all 15 factors are not truly critical to the specific task under investigation? In large scale experiments, we introduce candidate factors which rationally might be expected to be important to the task but may not be. Our first goal is to determine empirically which really are important. A screening study for 15 factors can be designed requiring as few as 32 and probably not more than 50 observations to provide the data needed to order the factors according to the magnitude of their effect on the performance of the specific task. The extra 18 observations are used to isolate critical two-factor interactions. It is unlikely that all 15 factors will be important; in fact a good guess would be that fewer than half will have large practical effects. In any case, even if only a few were eliminated by this screening process we have reduced still further the magnitude of the data collection process require to map the response surface. Furthermore, the data required to develop the higher-order response surface (if a test indicates it exists) are added in orthogonal blocks to the data from the screening study, a savings which helps keep the data collection economical.

Up to this point, nothing has been mentioned about the expensive habit of replicating complete designs. Still further economy is incurred when a multifactor study is performed by replicating only when it is necessary. In an earlier section of this report, it was shown how "hidden" replication provides adequate precision at considerable savings in data collection. The existence of trivial factors also provides an internal source of degrees of freedom for estimating an error variance if such is required. Finally, techniques of partial replication can be employed -- only selected conditions are repeated -- for an external estimate of error with which confidence limits of the response surface can be calculated.

One final word about design economy. Because fewer data points are collected, some information will be lost, presumably only the information the experimenter has determined is unimportant to the task. Still, in the absence of replication overkill found in traditional few-factor studies, the opportunities for bias to creep into the experiment are higher. Outside of careless data collection and a failure to control irrelevant sources of variance, the most common experiment-induced source of bias in psychological experiments comes from the need to collect data sequentially. Trend and trial-to-trial transfer effects are commonly found as a result of equipment drift and operator learning. In certain screening designs, there is a built-in protection against trend effects that requires no additional data collection (as is needed in traditional experiments employing counter-balanced designs). As a result, one run-through of a single design is sufficent to isolate any trend effects from the effects of interest. Additional data are required when trial to trial transfer effects occur or are anticipated. If subject characteristics critical to the task are treated as factors in the multifactor study, then in many instances total design replication, to account for those "individual differences," is unnecessary.

SECTION III

AWAVS PERFORMANCE SIMULATOR EXPERIMENTS

When the AWAVS carrier landing simulator is made available for research, current plans are to perform a number of performance experiments before actual transfer experiments are conducted (refer to page 7 for definitions). A brief description of several of these are given here.

GROSS-EFFECT STUDY

A preliminary comparison would be made of carrier landing performance by high and low skill/experience pilots on the "best" and the "worst" configurations of the simulator and under two levels of task difficulty. "Best" and "worst" in this case refer to the quality of the physical system, particularly the visual scene and the motion system.

This would serve several purposes. The information obtained, i.e., the differences in performance under the best and worst available simulator configurations, could influence future research plans. For example, if the differences are quite small then one may reconsider conducting the full scale multifactor study to identify only subtle effects of little practical importance. While this single experiment would not be sufficient to abandon all research, a small practical difference between best and worst conditions would certainly require the investigators to reevaluate their goals and priorities. If the difference between performances on the two simulator conditions is large, then support for a multifactor program is enhanced and the time invested in the preliminary effort has not been wasted. For example, it will have provided a means of trying out the equipment and the experimental personnel. It would have enabled the software, particularly that associated with measures of performance, to be fully developed and evaluated. It gives a chance for the procedures on running the study to be smoothed. All of these would be done under less demanding circumstances than would be found in a full-scale screening study.

INITIAL OVERALL SCREENING EXPERIMENT

A screening experiment will be conducted to assess the effects of approximately 13 factors associated with the visual and motion systems, the task, and the pilot, on pilot performance effectiveness in a simulated carrier-landing mission.

The candidate variables currently being considered for inclusion in the first experiment are:

a. Image quality (MTF): Carrier

b. Image quality (MTF): Seascape

c. Image quality (high-light brightness): Seascape

d. FLOLS systems

e. Field of view: Seascape

f. Velocity cues: Seascape x-y motion

g. Altitude cues: Seascape z motion

h. Platform motion

i. G-seat motion

j. LSO assistance

k. Task difficulty, turbulence

l. Task difficulty, A/C weight

m. Pilot carrier-landing experience

This experiment has been discussed in some detail in earlier papers (Simon, Vreuls, et al., 1977; Naval Training Equipment Center, 1976). A fictitious example of how it would be handled is described in Section V of this report.

VISUAL SYSTEM SCREENING EXPERIMENT

Because of the importance of the visual system in the AWAVS program, other experiments should follow the initial multifactor experiment; for example, content of the visual scene would be evaluated. Clues obtained from the initial screening experiment can indicate which visual scene variables that were studied are the most important. It can also indicate which conditions of the motion system are likely to affect design considerations of the visual scene. But there is a need for a more detailed examination of the visual scene, particularly in regard to content. The screening paradigm lends itself particularly to such a study, namely ability to study the effect on performance when certain objects, details, and informational clues are present or absent in the visual scene, as well as when certain physical parameters that affect

picture quality are set at less realistic levels. Given a large number of such variables, the screening study will permit them to be ordered according to their effect on performance in the carrier-landing mission. Later if considered necessary, for the quantitative variables, a more precise estimate of the function relating them to performance can be obtained with relatively little additional data collection.

SECTION IV

REFINING ECONOMICAL MULTIFACTOR DESIGNS

While the basic multifactor approach is well understood and is unquestionably the most informative and economical method by which controlled experiments of this type can be performed, there is still the need to refine and enhance its applicability to behavioral research. This is necessary since it was originally developed for use in other disciplines -- chemistry, agriculture, biology -- and may not always fit directly the peculiarities of behavioral research. Individual techniques employed in this approach to handle one aspect or the other of the experimental process may, in some cases, be combined to further improve their total capability. During this period of the contract, a number of techniques believed potentially relevant to the AWAVS program were investigated. (Note: understanding this section requires some background knowledge. See Simon, 1972, 1973, 1974, 1977a, 1977b).

FROM RESOLUTION IV TO V DESIGNS ECONOMICALLY

Screening designs are fractional factorials, generally of Resolution IV. This classification means that enough data will be collected to permit all main effects to be isolated from one another and from all two factor interaction effects However, the two factors interaction effects are not all isolated from one another; instead they are aliased in groups of independent strings.

Once the critical factors have been identified in the screening study, the investigator may wish to derive an equation in the form of a polynomial that approximates the response surface of proper degree. He will not want to start a new experiment; instead the economical approach would be to supplement the data from the screening study until at least a second order or higher order surface can be approximated. The classical central composite design is one popular data collection pattern for approximating response surfaces. The primary structure for this design is the fractional factorial, Resolution V. A design of that resolution is capable of isolating all main and all two factor interactions from one another. Thus, there is a gap between the size of the fractional factorial of the screening design at the end of the factor identification phase, and that of the fractional factorial at the beginning of the response surface phase. The question is: What is the most economical method of collecting the data required to fill this gap?

There were a number of papers in the statistical literature that had appeared potentially useful for solving this problem. The following represent some of the papers that were reviewed:

Draper, N. R. and Mitchell, T. J., Construction of the set of 256-run designs of resolution $\geq$ 5 and the set of even 512-run designs of resolution $\geq$ 6 with special reference to the unique saturated designs. The Annals of Mathematical Statistics, 1968, 39, 246-255.

John, P. W. M., Augmenting 2^{n-1} designs. Technometrics, 1966, 8, 469-480.

Pajak, T. F. and Addelman, S., Minimum full sequences of 2^{n-m} resolution III plans. J. Royal Stat. Soc., Series B, 1975, 37, 88-95.

Whitwell, J. C. and Morbey, G. K., Reduced designs of resolution five. Technometrics, 1961, 3, 459-477.

Addelman, S., Symmetrical and asymmetrical fractional factorial plans. Technometrics, 1962, 4, 47-57.

Addelman, S., Sequences of two-level fractional factorial plans. Technometrics, 1969, 11, 477-509.

Each represented some form of sequential approach to the Resolution V design through a series of blocks in which more sources of variance were isolated as more blocks of data were collected. The economy of this approach lay in the fact that the investigator could stop the data collection when all critical sources of variance had been identified.

After examining these and other papers, it was decided they offered no solution for the immediate problem since the initial blocks were not always the same as those used in the screening designs to be used in AWAVS, and when preplanned blocks are used more knowledge is assumed than is ordinarily available. They may result in unnecessary data collection. While other uses might be found for these techniques, it was decided that for the AWAVS problem, individual isolation of critical sources still seemed to be the best approach. This means that for any string of two factor interactions showing a critical overall effect, data would be collected to isolate which interactions accounted for the effect (Simon, 1973, pp 116-125; Daniel, 1962; 1976). Since the primary purpose in AWAVS is identification rather than response surface -- at least initially -- this procedure seems the most straightforward and least expensive. The same would hold true if there

is evidence that some three factor interactions might be present and biasing the main effects with which they are aliased. The individual, rationally guided search seems much more effective and economical than gross procedures for collecting blocks of data mechanically.

If the critical two factor interactions are isolated from the others, even though all two factor interactions have not been isolated, the result is for all practical purposes the same as if a complete Resolution V design had been used. This is referred to as a "reduced design of Resolution V."

SCREENING DESIGNS WITH SOME FACTORS HAVING MORE THAN TWO LEVELS

When screening designs involve qualitative factors, the investigator may wish to include more than two conditions of a particular factor. For example, in AWAVS there might have been three or even more distinct techniques for superimposing the ship scene on the background scene. Had this been the case, there would be no good basis for selecting which two should be used for the extreme cases needed in the screening design. Occasionally, even with quantitative factors, a design for handling a three level factor might be needed. There are times, for example, when a factor is not for all practical purposes continuous, and an investigator might wish to treat it as qualitative. More important are those factors that may show a total reversal in performance level over its range, sometimes referred to as a U-shaped performance curve. In that case, an investigator might wish to include a third level during the screening process rather than try to guess where the bend occurs in order to set one of the two levels at that point of maximum effect. How then might a three or four level factor be included in the conventional 2^{k-p} fractional factorial used as a screening design?

One might make the three level factor completely orthogonal to the other factors in the screening design. That would mean that the fractional factorial would be repeated three times, once each combined with a different level. While this is a clean approach, it might prove to be uneconomical. It would be more so if there were four conditions in the qualitative factor.

There already exist mixed level 2^m3^n and 2^m4^n fractional designs that have been published. However, these are usually limited to Resolution V fractional factorials which would be too costly to use for screening purposes.

Still a third technique is to modify the screening design to include a three (or four) level factor. This can be done economically by applying the Principle of Proportional Frequencies to the 2^{k-p} design. This principle states that

". . . a necessary and sufficient condition that the main effects estimates of two factors will be uncorrelated is that the levels of one factor occur with each of the levels of the other factor with proportional frequencies." Furthermore, it also states that ". . . for main effects to be orthogonal to two factor interaction effects, each combination of the levels of two factors must occur with the levels of another main effect with proportional frequencies."

Employing this principle, Addelman (1963, p. 60) shows how three two-level factors can be replaced by one four-level factor. Then he shows how a four level factor can be collapsed to form a three level factor, employing the same principle. Neither method is difficult to understand nor to do and so the details will not be repeated here. Because three of the two-level factors in the screening design must be sacrificed to include a three or a four level factor in the new design, the number of factors that can be screened in this modified design is reduced. There are times, therefore, when the size of the screening design would have to be increased to handle the desired number of factors.

If trend robust screening designs are used, the three or four level factors will not be as robust to trends as the individual original factors. Some combinations, however, are better than others and must be discovered for each design.

SECTION V

APPLYING ECONOMICAL MULTIFACTOR DESIGNS TO AMAVS PERFORMANCE EXPERIMENT -- AN EXAMPLE

In this section a fictitious example will be used to show how economical multifactor designs might be applied to an AMAVS performance experiment.

To reach this phase of the research program, it is assumed that the equipment has been built and debugged, both experimenters and pilot subjects have been properly and adequately briefed, the list of candidate factors has been chosen by experts after an informed analysis, appropriate performance measures have been selected, and the hardware and software required to obtain and analyze the information, either on-line or shortly thereafter, have been checked out. It is also assumed that this behavioral study is a dedicated one, that is, all who are involved with it have set as a primary goal the collection of information that will be of practical and enduring value.

EXPERIMENTAL OBJECTIVES

The experiment will have two primary objectives: one, to determine which of a large list of candidate factors supplied by experts have non-trivial effects on pilot performance for the specific task in the simulator; two, to obtain a response surface that describes the relationship between pilot performance and the simulator parameters for the specific task.

EXPERIMENTAL FACTORS

A list of candidate factors proposed for the first major AMAVS experiment on daytime carrier landing include the following: nine simulator factors, three task difficulty (environment) factors, and one pilot experience factor. These are listed on page 29 of this report.

Each factor will be studied initially at two levels, or two conditions. The levels would be set at practical limits of the operational space. The two conditions might be each of two alternatives, selected to represent the maximum range of difficulty, or they might be the presence and absence of some simulator characteristic. Subsequently other levels could be added if they exist and if the addition is warranted from an interpretation of the data already collected.

EXPERIMENTAL SUBJECTS

Pilots already capable of flying the simulator with minimum training would be employed in the first experiment. This is a performance study, not a transfer of training study. Two groups

with distinct skill/experience levels would be used. One would have practically no carrier landing experience; the other would have had considerable carrier landing experience.

EXPERIMENTAL PLAN AND PROCEDURE

The first step of the program is to identify which are the critical factors in the long candidate list. The strategy here is to avoid wasting time and effort collecting data about factors that have incidental or no effect on the particular task. Factors are included in the candidate list because they are believed to play a role in the general problem class; but only the experiment can determine to what extent each plays a role for the specific task under investigation. By quickly and inexpensively eliminating the factors of little practical importance, we can get on with the business of understanding the effects of the critical factors.

The identification process can best be achieved through the use of a "screening" design (Simon, 1975, 1977a, 1977b). There are several types of screening plans that might be selected depending on the availability of subjects and whether we intend to test each subject on all experimental conditions or not. It is impossible to discuss here all of the alternatives that must be considered by the experimenter and the nuances involved in selecting one or the other. There is no cookbook approach; the experimenter must be knowledgeable about what to consider, the alternatives available, and the consequences of each decision. We will, by way of illustration, select a particular design that would permit us to test a pilot on all experimental conditions without concern for the more common trend effects -- linear, quadratic, and cubic -- that might bias the effects of interest. If skilled pilots are used and precautions taken to minimize trial-to-trial carry-over effects, as an initial effort, such a study can provide an immediate overview of the problem and provide clues as to what the next step should be*.

The data collection plan would be a Resolution IV design of the form shown in Table 4 that is capable of estimating the main effects of up to 16 factors independently of two factor interactions by testing performance on 32 experimental conditions. The special feature of this particular screening plan is that the experimental effects, e.g., of the simulator and the task difficulty factors, will be minimally biased if there

---

*An alternate approach would be to run a different subject on each experimental condition (i.e., equipment configuration).

"GOT" DATA	ENVIRONMENTAL CONDITION	MAIN EFFECTS													THREE FACTOR INTERACTION DEVIATIONS			
		(1)	A	B	C	D	E	F	G	H	I	J	K	L	REF	MAX	MIN	RANG
1	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
42	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
46	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
47	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
48	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
49	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
50	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
51	[illegible]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
REGENT	LINEAR																	0
TREVORROW	SUBSTITUTE													0	0	0	0	
OVERLAP	CARDS							0	0	0	0	0	0					
PRETTY LONG CHANGE COUNT		0	0	20	22	18	16	23	19	17	27	25	25	18	24	26	30	31

MAIN EFFECTS ARE ALLOWED WITH THREE FACTOR INTERACTION ERRORS
ONLY ONE THREE-FACTOR INTERACTION IS THE ERROR IS SHOWN HERE
ONLY ONE TWO-FACTOR INTERACTION IS THE ERROR IS SHOWN HERE
REMAIN SPACES REPRESENT ZERO ERRORS. TERMS WITH ERRORS IN THEIR REPRESENTS SOME ERRORS

[illegible]

[illegible]

[illegible]

[illegible]

are linear, quadratic, or cubic trend effects (e.g., subject learning, equipment drift) running through the data. This particular Resolution IV screening design is said to be robust, or resistant, to trends (Simon, 1977a).

Pilot experience may be treated as any other factor and included within the experimental design, or as in this example, may be introduced as an additional factor, outside of and orthogonal to the design. The decision to include an experimental factor within or outside the design -- for example, task difficulty factors might also be added outside the screening design -- depends on logistical considerations balanced against economy and information quality. In our example, we will keep the twelve factors within the screening design and pilot experience outside it. Thus, in this example, each pilot used will be tested on all 32 conditions, and at least one high and one low experience pilot would be studied.

Before continuing with the description of the experiment, let us examine the characteristics of this particular 2^{12-7}_{IV} screening design (see Table 4). There are 32 different experimental conditions purposely selected out of a possible $2^{12} = 4096$ in the complete factorial. Each row of the experimental design represents a different experimental condition. The plus or minus sign in the column under each factor (main effects only) shows which of the two levels the experimenter would use when setting up each condition. Conditions are to be run in the order shown.

The considerations involved in handling multiple performance measures, the dependent variables, are much too complicated to discuss here. Therefore, for this example, we will assume that a decision has been made and for each condition a single or composite performance score has been obtained. The experimental conditions are selected so that we base our estimate of the mean of each condition of each factor on 32 observations. We can estimate the main effect of each factor independently of one another and of any two factor interaction. Each mean, however, will be aliased with a string of three factor interactions. The effects of still higher-order interactions are also aliased with these effects but can be ignored since the probability that they would have any practical effect is negligible. Since the design is capable of handling up to 16 factors and we will use only twelve columns, the design provides some information regarding strings of three factor interactions not aliased with main effects. The effects of strings of two factor interactions with eight or fewer different interactions per string can be estimated independently of one another and of the main and three factor interaction effects. This data provides clues regarding the presence of critical two factor interactions. In the screening phase, knowledge of interactions is only important if it affects the selection or elimination of a factor.

A prime feature of this design is the order in which the experimental conditions are presented to the subject for testing. They are ordered in the design so that when the data is collected, no main effect will be biased by any linear or quadratic trend running through the data and only two would be affected trivially (1 or 2 percent) by a cubic trend. The actual values are shown below the experimental design in Table 4. This is an important advantage when a single subject is tested serially. The resistance to trend occurs with this design without having to reduce the economy of the design by adding more conditions or counterbalancing the ones that are used.

If changing the level of a factor is difficult or time consuming, then the proposed experimental design per se is cumbersome. In the AWAVS experiment, changing the circuit boards for the MTF of the carrier image may become very time consuming since the equipment must be turned off during the change and then warmed up after it has been turned on again; delay can disrupt a subject's rapport. Several methods are available to handle this situation. One, the particular factor could be pulled outside the design and changed only a few times while the remaining factors are nested within it. Two, the design shown in Table 4 can be modified in a way that will reduce the number of changes required. In making this modification, however, the degree to which the design is resistant to trends is diminished slightly (Simon, 1977a). Three, the best method, when feasible, is always to modify the equipment to simplify changing conditions. While possibly initially costly, for any extended research program, it can be justified by the savings in time and the improvement in data quality.

Analysis of the First Set of Data

Once the performances at the 32 data points have been measured for a single pilot, whatever his experience level, the data can be analyzed. This analysis is extremely simple, consisting of finding the mean difference between high (+) and low (-) conditions in each column. This can be expedited by using Yates' algorithm (Simon, 1977a).

The results of such an analysis is illustrated (using fictitious data*) in Table 5. In this example the twelve

---

*The numbers were taken from an actual experiment, so they do reflect what can be expected from a real experiment. However, the context in which they appear has been modified to fit the example.

TABLE 5. ANALYSIS OF FICTITIOUS AWAVS DATA FROM DESIGN IN TABLE 4.

Rank (largest 1st)	Source	1 Mean Difference (Effect)	2 Eta Squared (η^2)	3 Cumulative Proportion of Variance Accounted For
31	E	.3359	.2662	.2662
30	A	.2422	.1384	.4046
29	G	.2266	.1212	.5258
28	(AEF,...)*	- .2266	.1212	.6470
27	F	.1797	.0762	.7232
26	K	.1172	.0324	.7556
25	AF,BC,DL,GH,IJ	.1172	.0324	.7880
24	D	.1016	.0244	.8124
23	AJ,CE,FI,HK	- .1016	.0244	.8368
22	AK,DE,GI,HJ	- .1016	.0244	.8612
21	EL,FK,GJ,HI	.1016	.0244	.8856
20	AI,BE,FJ,GK	- .1016	.0244	.9100
19	I	.0859	.0174	.9274
18	BK,DI,EG,JL	.0703	.0116	.9390
17	AE,BI,CJ,DK	- .0703	.0116	.9506
16	(ABK,...)*	- .0547	.0070	.9576
15	H	.0547	.0070	.9646
14	(AEL,...)*	- .0547	.0070	.9716
13	AB,CF,DG,EI,HL	.0391	.0036	.9752
12	AC,BF,DH,EJ,GL	.0391	.0036	.9788
11	AH,BL,CD,FG,JK	- .0391	.0036	.9824
10	B	- .0234	.0013	.9837
9	J	.0234	.0013	.9850
8	(AEH,...)*	- .0234	.0013	.9863
7	BJ,CI,EF,KL	- .0234	.0013	.9876
6	AL,BH,CG,DF	.0234	.0013	.9889
5	CK,DJ,EH,IL	- .0234	.0013	.9902
4	AD,BG,CH,EK,FL	- .0234	.0013	.9915
3	C	.0078	.0001	.9916
2	L	.0078	.0001	.9917
1	AG,BD,CL,FH,IK	.0078	.0001	.9918

*Represents a string of three-factor interactions

simulator and task difficulty factors were included in the design and the problem of level changing has been solved without modifying the design. We will examine the fictitious results from a single pilot tested on all 32 conditions.

The results in Table 5 listed the effects of each source of variance -- main, two factor and three factor interaction strings -- in order of their magnitude (Col. 1) The proportion of the total variance contributed by each independent source is shown in Col. 2. The cumulative proportion accounting for all sources as each succeeding one is included is shown in Col. 3.

The investigator must decide which sources of variance are critical. Within some reasonable limits he can probably state what minimum size effect (difference) he considers to be of practical importance. He will ordinarily have little difficulty eliminating those very small effects that would be considered trivial. He can also recognize the obviously critical factors which have very large effects. Therefore, the major problem for the investigator is to decide which of the marginal effects are to be considered important. Let us say for this illustration that a mean difference (an effect) of less than .10 is probably trivial. That would mean that Factors E, A, and G are probably critical, while F and K are marginal for this particular task (and within the limits set by the experiment) and Factor D is right on the line*. If Col. 2 is examined, we can see that Factor F accounts for approximately eight percent of the variance in this experiment and Factor K accounts for three percent. The other three (E, A, and G) are markedly higher. If we examine Col. 3, we see that for main effects only, if Factors E, A, G, F, K, and D are terms in a first order polynomial, the regression would account for approximately 66 percent of the total variance.

If the effects of <u>all sources</u> up to and including Factor D were included in a regression equation, we would account for 81 percent of the total variance. If all sources up to and including Factor K were included in an equation which would be essentially a first order polynomial with an additional term representing a string of three factor interactions, we would account for 76 percent of the performance variance in this experiment. The 76 percent represents a multiple correlation of .87, which is respectable since it is

---

*There are other considerations that would be involved in this interpretation, too detailed to describe here. Once again, the investigator cannot analyze his data mechanically; he must understand the process and apply it wisely.

based on five factors out of twelve originally believed important by a group of experts, and in fact, represents a prediction based on all 12 -- for this task, subject type, and within the limits of the experimental conditions.

But we cannot arbitrarily add or dismiss sources of variance in this way. We could make ourselves "look good" by adding more and more, although it would have little meaning operationally. We need other criteria to make our selection at the point the differences approach the trivial level and, the proportion of variance accounted for by each new addition is small. Although there were no replications in the design by which to estimate an error variance (this will be further discussed later), we can use order statistics to estimate what the error variance is and whether an observed effect is larger than one might expect to find by chance.

In Figure 2, a half-normal plot is shown of all 31 effects -- the mean differences -- of the study. The slant line represents a normal distribution of a set of effects. All effects located to the right of this line would therefore be considered larger than one might expect by chance. It is clear that neither the effects of D nor K in this study were larger than might have been expected by chance. The four factors E, A, G, and F, along with the string of triple interactions, accounted for 72 percent of the variance, yielding a multiple correlation of .85 $\pm$ .10.

The study would be repeated using the pilots with different amounts of experience. Examination of both sets of results, separately and in combination, looking for patterns and for marked differences, would be an important part of the analysis.

Having reached this point, an investigator has a number of choices. If the only purpose of the experiment is to identify the critical factors, we have come close to it already. Whether or not Factor K or any of those with even smaller effects would be used at the level (configuration) producing the highest performance is no longer a decision based on performance. Since the differences in performance are marginal, costs and technical considerations become the overriding criteria. In a program such as AWAVS, other criteria, e.g., transfer effectiveness, can also determine which configuration would be used.

The first objective of this experiment has still not been met until we have answered a few more questions. One of them is: What interaction(s) within the string showing the large composite effect actually accounted for that effect? It is possible that that interaction might include a factor that was not one of the four selected as critical. In other words, before we can be sure we have not omitted a critical factor, we should collect some additional data to see which one of the triple interactions in the string (listed in Table 6) was responsible for the large effect.

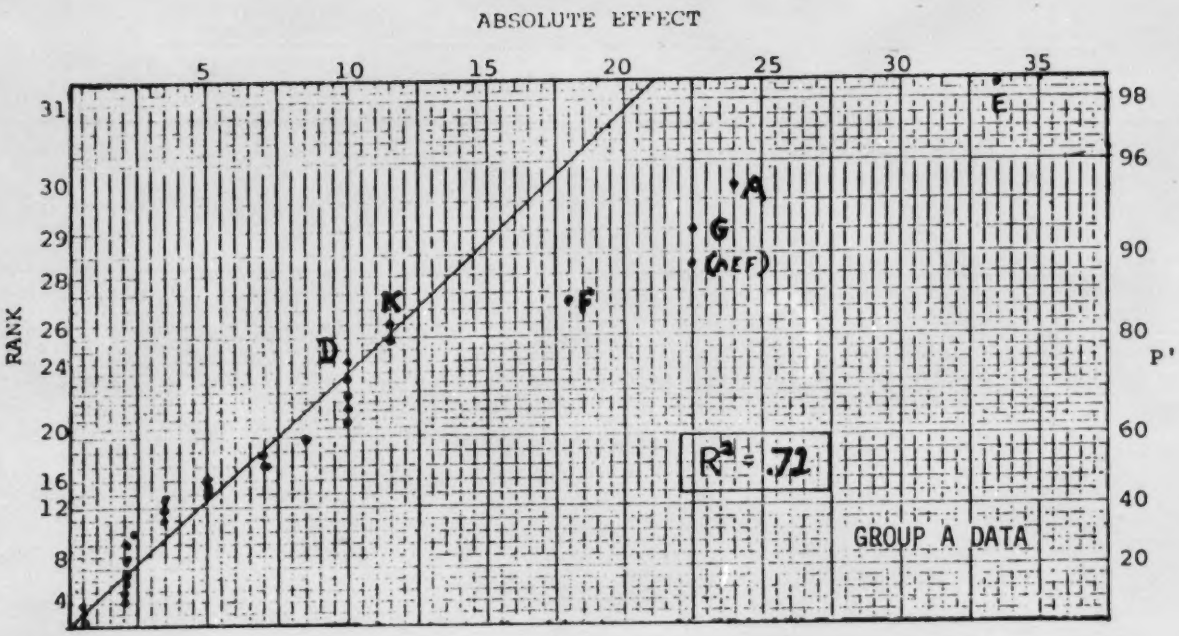

Figure 2. Half Normal Plot of Experimental Data in Table 5.

TABLE 6. THREE FACTOR INTERACTIONS IN THE CRITICAL STRING IN TABLE 5

ABJ	ACI	AEF	BCE	CEF
CGK	DGJ	DHI	EGH	EIJ

If we wish to isolate the effects of each of these interactions from one another, we would have to collect performance data at a minimum of ten new coordinates, although for the sake of balance, 16 would probably be used. However, we can make some preliminary guesses that might reduce the effort. For example, if we only considered the interactions that were composed of some of the four factors that we knew were critical, we would only have to isolate

AEF.

This is also the one identified if we were to consider those containing all factors in the upper half of the plot. In this way, if we find it does account for most of the observed effect in that string, we'd not have to collect any more data. In theory, we could estimate the effect of the AEF interaction in the same way we estimated the effect of Factor A, by finding two conditions, one of which represents the + condition of Interaction AEF and one which represents the - condition of AEF. Obviously, conditions aef and (1) would serve these requirements. Also abef and b, acef and c, abcdef and bcd, and so forth. Several of these might be used to increase the reliability of the estimate.

If the magnitude of the AEF effect did not correspond with that found in the study -- and one must allow some leeway for differences in the data collection process -- then one must look further and begin to suspect that the critical interaction is a disordinal one. In this example, however, it would be highly unlikely that this were the case, but if it were necessary to isolate the remaining sources, a balanced plan (see Simon, 1973, pp 120-123) might be employed.

The chances are good that this quick approach will work since most interactions found in the behavioral sciences are of the ordinal type. In that case, the large interactions would be associated with the large effects and can be eliminated by rescaling the dependent variable. The less frequent disordinal type of interaction is the more important one, with which the interaction may be large while the main effects making up those interactions might appear trivial. Since these interactions cannot be eliminated by some transformation of the data,

they are sometimes referred to as "intrinsic" interactions. If we wish to be certain that we have found all critical main effects, we must be certain we have detected any that contribute to disordinal interactions.

In this example, no strings of two factor interactions were found to have critical effects, although the one set (located in rank between Factors K and D) might be a possible candidate. Ordinarily, there is a greater chance of having a critical two factor effect than a three factor effect. It is interesting to note that although this string did not show a large effect, interaction AF was in the string. With Factors A and F and interaction AEF all large, it is not surprising that the string with AF was also large; however, inspection of the half-normal plot (Figure 2) suggests that an effect of this magnitude would probably have occurred by chance. Whether in fact it did account for the proportion of variance shown in the string would have to be tested by the addition of new experimental conditions as was done in the case of the three factor interaction.

There is one point that should be remembered in regard to strings of interactions: it is possible for two large effects to cancel one another. While the chances are not necessarily high, the investigator must be alert for that possibility. The analysis that should precede an experimental effort will often supply the investigator with the cues necessary to anticipate this situation.

At this point in the investigation, we should have identified all of the critical factors out of the candidate group, including those that might have been hidden within a disordinal interaction. The cost of such an effort, to study 12 equipment and environment factors plus pilot experience in the manner proposed, would be the costs of collecting data on 2 x 32 = 64 observations, plus possibly an additional twenty or so observations. Had we decided to make the subject factor a part of the Resolution IV design, then the study might have been concluded with as few as 50 observations. Certainly this is sufficient to obtain the information of any practical importance.

About the only weakness at this point is in the assumption that two experience levels are sufficient to classify the pilots, and that all pilots within these two groups would in fact be homogeneous. If they are, then our experiment, insofar as objective one is concerned, is complete. If they are not, it is not the design that is at fault, but the original planning, for the intent is to identify all critical factors including pilot characteristics that might influence simulator design.

More pilots may have to be run in the latter case, but not in a haphazard manner. Identification of the other pilot dimensions becomes a crucial issue, somewhat oblique to the original objective but one which could influence the interpretation of the results. In practice, it is highly unlikely that only one pilot of each type would have been run; still it is important that when more are included, it is because we wish to extract more information, not that we just wish to be redundant.

Obtaining a Response Surface

A response surface is merely a representation of the multidimensional functions relating performance to the critical experimental factors. It is frequently represented by a polynomial equation derived from the experimental data. While an equation can be written whether the factors are qualitative or quantitative, continuous or discrete, the concept of a response surface implies that the variables involved can be described along a continuum.

In the primary AWAVS study, as it has been planned, most of the factors are either qualitative or dichotomous and discrete quantitative factors, and as such, do not need to be represented by a response surface. For all practical purposes, the experiment would stop when all the critical factors had been identified and the best configuration identified. However, for purposes of illustration, we shall continue this section using the AWAV example to illustrate the steps involved if we wished to approximate the best fit of a response surface were the variables of the appropriate type.

The data from the screening design can be used to write an equation containing only linear terms:

$$\bar{Y} = .543 + .168\ E + .121\ A + .113\ G + .090\ F - .113\ AEF$$

with each coefficient equal to one-half the mean difference for the corresponding effect. The interaction AEF is a linear interaction, i.e., linear A x linear E x linear F. Before final acceptance, the residuals from this equation should be analyzed (Daniel, 1976).

If an investigator plans to develop a response surface, he should include center points in his experimental design during the screening phase. These center points are at coordinates (0, 0, 0, ...0, 0) in the center of the incomplete hypercube defined by the 2_{IV}^{12-7} fractional factorial. Several measures at the center would be taken, preferably at equal intervals along the 32 condition run. Since we will continue this example and assume that the 12 factors were in fact quantitative and continuous, we will have already included center point measurements of performance at the beginning and end of the 32 condition run and after the 8th, 16th, and 24th conditions, making a total of five center points in all.

At this point, we do not know whether the linear equation shown above adequately represents the true response surface. It is not uncommon to find the relationship between performance and factors in behavioral studies to be non-linear. The center point data provides us with an opportunity to test to see whether there is curvilinearity in the response surface, for if all dimensions were collapsed onto a single dimension, we would have measures at three levels of each factor, enough to test to see if a quadratic relationship would better describe the data. If a Lack of Fit Test reveals that the linear equation is not adequate, then the investigator must be prepared to collect more data.

His first goal is to collect enough data to write a second degree polynomial, which would include all critical main effects, all critical two factor interactions, and all critical quadratic terms. In this study, we have already determined that linear two factor interactions have probably only trivial effects and that there is one important linear triple interaction and so in that regard, we are ahead of the game. Still we will want to add some points to estimate the quadratic terms. One data collection plan for this purpose is called a "central-composite" design (Simon, 1970b, 1973, 1976a, 1977a,b).

The classic central-composite design is composed of a 2^{k-p} Resolution V factorial hypercube, a 2k star portion, and some center points, where k equals the number of factors and p is the fraction of the complete factorial needed to satisfy the requirements of a Resolution V design. With that design all main effects and all two factor interactions would be isolated from one another. The screening design, already completed, provided us with a 2^{12-7} Resolution IV design in which all main effects were estimated independently of one another and of the two factor interactions, but within sets of independent strings, two factor interactions were still aliased with one another. Ordinarily the investigator might collect more data to make the Resolution IV design a Resolution V design, or he may find another solution that does not require more data. There is such a solution in this example.

From the results of the screening study, it had been concluded that only four factors were critical. If we were to drop all letters representing the non-critical factors from a completed design in which all aliased two factor interactions are shown, i.e., Table V, we would find that the original 2^{12-7} Resolution IV design becomes, for all practical purposes, a Resolution V+ design. Had the three factor interactions in the strings been listed, it would have been seen that effects of a complete 2^4 factorial are estimated since all other effects were judged trivial. Note that the six possible two factor interaction terms for the four critical factors are all estimated independently of one another at ranks 25 (AF), 18 (BG), 17(AE), 11(FG), 7 (EF), and 1 (AG). The effects of these

in this experiment were judged to be inconsequential. Therefore, although no more data has been collected, we have, for all practical purposes, the Resolution V design required for the fractional hypercube portion of the central-composite design.

In fact, if as a precaution in writing the response surface, the investigator preferred to include Factor E in the equation, albeit marginal, the existing data is still sufficient to estimate the ten two-factor interactions for these five factors, all independent of one another. The additional two factor interactions can be found at ranks 22 (AE), 21 (DE), 20 (CE), and 4 (BE). The remaining variances at ranks 24, 23, 19, 16, 15, 14 13, 12, 10, 9, 8, 6, 5, 1, and 2 would be combined to make up the "error" variance*.

If we perform a Lack of Fit test -- using the center points for this purpose -- and find that a test of the linear fit is poor, then data should be collected at the "star" points to estimate the coefficients of the quadratic terms for the five factors. These points are located at coordinates ($\pm\alpha$, 0,0,0,0,), (0, $\pm\alpha$, 0,0,0,), ... (0,0,0,0, $\pm\alpha$). The value of α depends on other features of the design, and a discussion of how it is selected is too involved for this paper. The central-composite design requires that the number of star points equal two times the number of factors in the experiment. Therefore, if the investigator decides to keep the five factors, he must collect data at a minimum of 2 x 5 = 10 additional points. When the star points are combined with the points of the fractional hypercube and the center points in the screening design, five measurements will have been made along the scale for each factor. While this does not produce a 5^5 factorial design, the points are located so that estimates of the quadratic terms can be obtained.

Since we presumably had identified all critical two and three factor interactions during the screening phase by collecting data at a total of 32 (cube) plus 5 (center) plus 10 (star) equals 47 experimental conditions, we have approximated the response surface for a five factor space. However, it should be remembered that we began with a 12 factor space of which only the five had critical effects in the particular task. If the 12 factors originally selected by the experts were in fact the most likely candidates influencing performance on the task under investigation, then this laboratory-derived equation of the only truly critical five out of 12 factors should be

---

*Actually these contained the higher-interaction terms required to complete the 2^5 factorial -- all shown to be negligible.

expected to predict performance under operational conditions quite well. A different response surface would be derived in the same way for each pilot experience level investigated.

VERIFICATION AND FIDUCIAL LIMITS

Once an adequate equation has been derived, depending on the time and resources available, the investigator may wish to do two things: 1) to establish confidence limits, and 2) to verify the equation. The first might be done by replicating the existing design at select points -- a partial replication. The second might be done by selecting combinations of factors where no previous data had been taken to see if the equation would predict the results within acceptable confidence limits. The real test for verifying the equation would be to collect data under field conditions to determine how closely the equation would predict it. Unmentioned in the above discussion, but critical in any holistic approach to a problem, is the handling of uncontrollable variables. If they can't be manipulated, then they should be measured and their effects isolated from the other data through some covariance analysis.

SECTION VI

QUASI-TRANSFER EXPERIMENTS

There has actually been very little research seeking fundamental principles of transfer for the pilot training situation. Many studies have been conducted for the purpose of evaluating existing devices and as such do not provide the information needed to optimize design. Some studies performed with the intended purpose of answering fundamental questions have been so narrow in their context that it would be foolhardy to generalize beyond the conditions of the particular experiment. Extrapolations from the results of classical transfer of training studies -- often on verbal material or oversimplified perceptual-motor tasks -- cannot be made with confidence, at least insofar as recommendations regarding specific design decisions are concerned. It is therefore desirable to pursue studies in the context of pilot training simulators such as AWAVS that seek principles of transfer of training. For this purpose, quasi-transfer experiments can be considered as an economical but effective approach to use.

A "quasi-transfer" experiment for the AWAVS program is defined as one in which performance is never measured under realistic, i.e., non-simulation, conditions. For pilot training this means that the experiment would include no post-training periods in which performance would be measured in the aircraft. Instead, an alternate simulation configuration would be used to represent the flight conditions.

This artificiality makes it necessary to interpret experimental results with caution. They may be used to understand the transfer of training process, but should not be the basis -- without considerable experience and support data -- for evaluating the transfer of training qualities of the AWAVS simulator. Whatever differences exist between the simulator configuration representing the aircraft and the actual aircraft -- differences that may not be evident to the investigator -- could seriously distort interpretations regarding transfer from the simulation experiment to the specific aircraft. These considerations, however, should not discourage use of a simulator to understand conditions affecting transfer of training in general. In summa, we would use the quasi-transfer experiment to discover what transfer of a particular nature, quantity, and direction (i.e., positive or negative) would be affected by specific simulator characteristics. Understanding these things in depth would facilitate our ability to make better design decisions in future simulation efforts, and help us to plan and conduct real transfer studies.

FIDELITY

No single unproven principle dominates the design of pilot training simulators more than the "fidelity principle." This principle implies that:

> Transfer of training from simulator to aircraft is a positive function of the degree to which the simulator faithfully reflects the characteristics of the aircraft.

In Figures 1A through 3C, graphic representations of this principle along with cost considerations have been reproduced from several reports on this topic (Hicks, 1970; Roscoe, 1970). That "fidelity" has never been adequately defined has not deterred the use of this principle which has its roots in classical psychology studies of transfer. For some, fidelity implies physical realism; for others it suggests that psychological similarity is probably more important. On the other hand, some such as Cara (1973) believe that how the simulator is used is more important for optimizing transfer than the degree of simulation realism.

Evidence that realism is important is attributed from applications of simulator training, as employed by the commercial airlines to train and upgrade pilot skills. There have also been component studies (often under simplified conditions) that purport to demonstrate the validity of the principle. Other component studies purport to demonstrate that the principle does not hold. However valid the fidelity principle may be, costs and state-of-the-art of simulation place considerable pressure on those who design the simulators to move as far away from a faithful reproduction of reality as is compatible with effective training. In spite of large outlays of money for research, no experiment to date has provided definitive answers nor has been sufficient to specify those conditions under which fidelity is required nor to dimensionalize fidelity into its composite parts and demonstrate the conditions under which each component is important to transfer.

Dimensionalizing the Situation

Before the fidelity problem can be attacked properly, the situation in which fidelity is to be examined must be more thoroughly dimensionalized than it has been in the past. While most of these characteristics have been recognized in discussions of fidelity, few investigators seem to see the need to specify the [illegible] of the multidimensional space which their experiment is intended to illuminate. Human behavior is situation specific. To discuss "fidelity," we must discuss it in the context of a situation. Dimensions of an NAVTRAEQUIPCEN situation include:

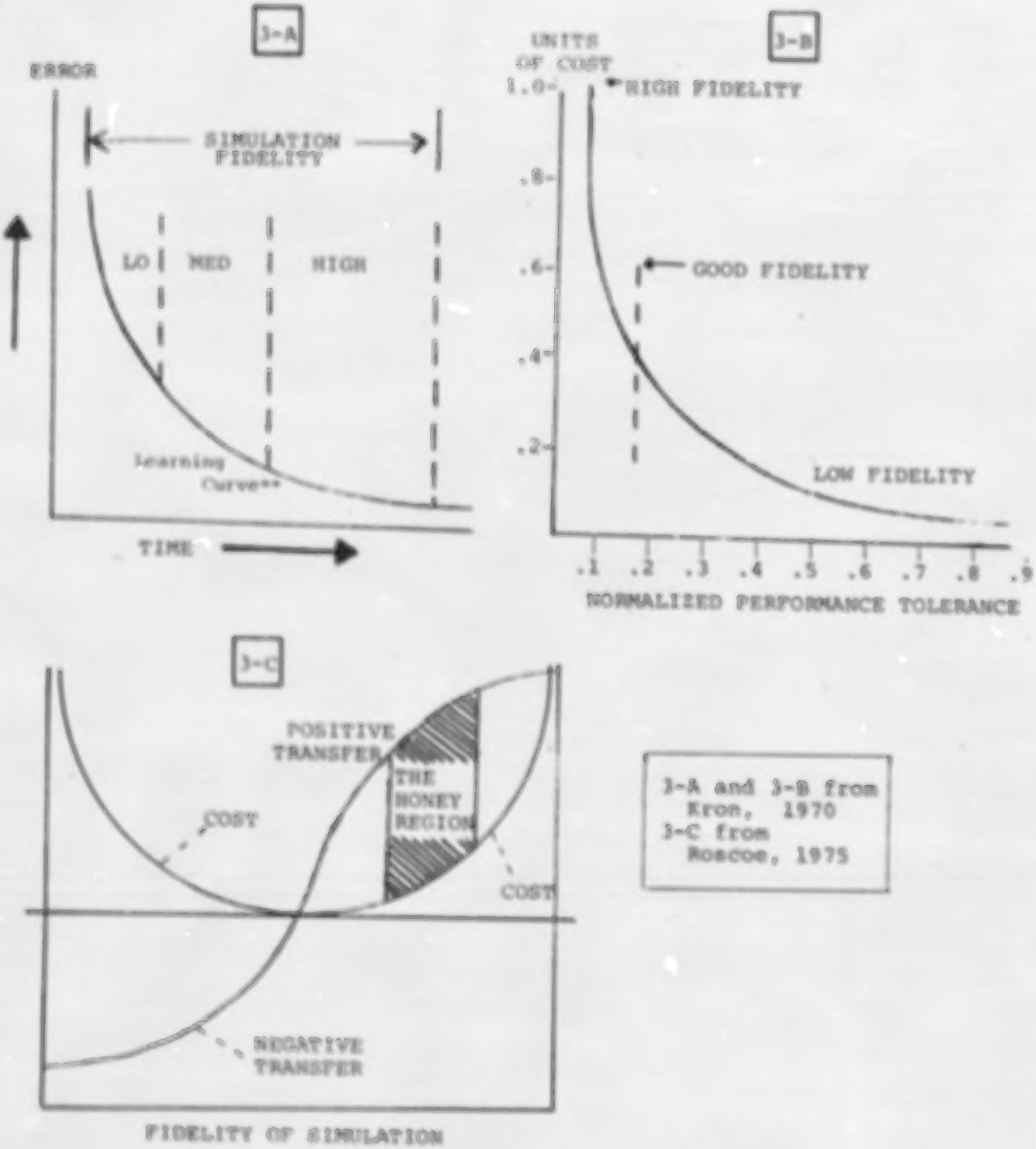

Figure 3. Theoretical Relationships Between Transfer, Simulation Fidelity and Costs

Pilot skill

Pilot experience

Task complexity

Simulator mission

Simulator complexity (i.e., aircraft simulated)

Simulator component (e.g., visual system)

Other critical considerations include:

Training curriculae

Instructor skill

Performance criteria

Dimensionalizing Fidelity

Simulation fidelity has generally been evaluated in terms of known similarities between physical systems or on the basis of pilot judgment. There have been proposals wherein performance equivalence on a simulator and aircraft would be interpreted as perceptual equivalence, implying a measure of effective fidelity. But these approaches have two weaknesses: 1) they presume that fidelity is a single entity and simulator fidelity becomes a gross measure; 2) they don't answer whether or not faithful simulation is a necessary feature at all. Certainly there are recognized examples where a simplification in some case or increased difficulties in others have been successfully employed to improve transfer of training. This implies that research in fidelity should break fidelity down into meaningful parts and to ask the more general question: Under what conditions are the components of fidelity important and under what conditions are they not in the training context?

Some examples of the more obviously different ways in which fidelity of the visual or motion simulation system can be dimensionalized are shown in Table 7.

Experiments

Given an appropriate simulator*, experimental questions relevant to an understanding of fidelity and its effect on transfer of training can be examined.

---

*What an "appropriate" simulator is will not be defined here. The answer is probably pragmatic -- it will be appropriate if it is available, has the degree of flexibility suitable for research, and represents the AWAVS type tasks.

TABLE 7. EXAMPLES OF DIMENSIONS OF FIDELITY IN VISUAL AND MOTION SIMULATION SYSTEMS

Type of Fidelity	Examples
Visual system:	
Continuous variables that may be decreased or increased	Resolution; brightness, contrast
Spatial distortion	Size, shape, patterns
Temporal distortion	Speed of response; lag relative to compatible motion system
Incompleteness	
Omission of objects	Realism of background content
Omission of detail	Sea texture
Skeletal, pictorial or symbolic	
Added information	Attention getters; emphasizers not found in real world
Motion system:	
Simplified model	Aircraft dynamics; omitted degrees of motion
Distorted feel	Aircraft dynamics; motion kinaesthetic cues

How do reductions in fidelity affect system transfer? A quasi-transfer study might be conducted using one simulator configuration to represent the real world, i.e., the aircraft, and all other configurations to represent varying degrees of reduced fidelity. If preliminary studies relating fidelity to performance were conducted first, an investigator might use that information in planning this study. Ordinarily the most sophisticated simulator configuration might be used to represent the aircraft: on the other hand, for certain classes of variables, no particular configuration need be singled out. Instead the study would be conducted to find out what happens to transfer when fidelity increases or decreases, when transfer is positive or negative as a function of the psychophysical characteristics of simulator components.

Experimental designs described earlier for economically performing large multifactor performance studies might be employed in these experiments. Subjects would be trained on simulator configurations differing in their fidelity and subsequently tested for transfer on another simulator configuration representing the aircraft. In addition to providing a comprehensive picture of the transfer problem in complex simulation and task situations, these studies would also provide a chance to experience, evaluate, and learn more about proposed economical transfer designs prior to their use under "real" conditions. Some experimental data collection plans, described later in Section VII, could be examined in a quasi-transfer study in order to improve our transfer of training research methodology.

Novel Transfer of Training Designs

Simon (1974) reviewed a class of experimental designs, called "change-over," "cross-over," "carry-over," or "residual" designs, that might make the study of transfer principles more economical if they were employed. Unlike the designs used in a conventional transfer experiment, these permit a single subject to be tested on a number of configurations serially, while being able to measure the residual effect carried over from one configuration to the one that follows it. The designs are capable of isolating the direct effect of the configuration being tested on the particular trial from the residual effect carried over -- transferred -- from practice on a different configuration used on the previous trial.

This class of design lends itself particularly to quasi-transfer experiments, where the simulator can be used for all the configurations under investigation. Each configuration will be preceded and followed by every other configuration, so that at the end of the experiment, we can determine which configuration has the largest average residual (transfer) effect on the performance of the configuration that followed it. If there are interactions between direct and residual effects so that

the amount of transfer due to one configuration depends on which particular configuration follows it, then this too can be analyzed, although the designs for this purpose are more complicated. By having the configurations in the series vary in more than one dimension, the relationships between simulator components, fidelity, and transfer may eventually be determined.

Measuring sequential transfer is not completely new to psychologists who have included "order" in designs. In those cases, with only two conditions, A and B, to be studied, half of the subjects are run on order A to B, and half on B to A, and the effects evaluated. Ordinarily this has been done for cleansing rather than for informative reasons.

Change-over designs appear in two basic forms: one requiring a number of subjects (where direct and residual effects are balanced across subjects) and the other in which estimates of residual effects are balanced within the responses made by a single subject tested serially.

For example, here is a design in which four experimental configurations that differ in their similarity to one another along a known dimension (or dimensions) might be used to determine the amount of transfer that can be attributed to conditions A, B, C, and D:

		Subjects			
		I	II	III	IV
Trial (Period)	1	A	B	C	D
	2	B+a	D+b	A+c	C+d
	3	C+b	A+d	D+a	B+c
	4	D+c	C+a	B+d	A+b

The capital letters indicate which experimental condition (A,B,C, or D) is being tested. It's effect is referred to as the "direct" effect. But performance in these serial presentations may also be affected by "residual" effects carried over from the previous configuration, as indicated by the small letters (a,b,c, and d). Performance as it is measured on any trial is the composite of both the direct and residual effect. The direct effects are distributed in the arrangement of a balanced Latin square with each condition preceding and following every other condition (vertically) once and only once, and also appearing once in each column and each row. Direct and residual effects can be independently estimated by adding a fifth trial (row) in which the conditions of the fourth row are repeated:

5	D+c	C+a	B+d	A+b

The total variance of this extra period design can be partitioned as follows:

Subjects
Trials (periods)
Direct effects
Residual effects
Error

There are several variations on this plan (see Simon, 1974). Its major limitation is that it assumes that the residual for any configuration (or condition) is constant irrespective of what configuration follows it. For the most part, these designs are not used factorially, that is, the four conditions ordinarily do not represent a 2^2 set of conditions, although there is no reason why they cannot.

Other designs are available when direct and residual effects are assumed to interact. However, these designs have never really been optimized, have seldom been used, and ordinarily increase the amount of data collection required. If we seriously wish to develop new economical methods of studying transfer, this class of design should not be overlooked.

A different type of design, referred to as a <u>serially balanced sequence design</u>, can be used with a single subject tested repeatedly on all experimental conditions. One example for four conditions is:

——————>

B; B C A D; D C B A; A B D C; C A D B;
B D A C; C D B A; A B C D; D A C B

Block effects, direct effects, residual effects, and error can be estimated with designs of this type although their effects are not always orthogonal. Sequences are usually balanced against direct and residual interaction effects although in the past these interactions have not been isolated. Both serially balanced and carry-over designs can be adapted to measure not only first residual, but second residual effects that occur two trials after the direct effects were introduced.

Where the effort can be made at relatively low costs, an attempt should be made to employ this class of design if for no other reason than to establish its value for the experimental study of transfer and simulator fidelity. If effective, it can represent a less expensive means of learning something quickly about transfer. It is apparent that these designs lend themselves to only certain problems, particularly where training to use the simulator has taken place prior to the experiment and suffices for all configurations. For designs in which the

residual effects are assumed to be additive to each direct effect, we would hope to find out which configuration is likely to result in the highest overall effect being carried over to the configurations that follow. The assumption is made that on a relative basis this would hold true were the real aircraft involved. On the other hand, if designs are used in which direct-by-residual interaction effects can be isolated, we may discover more fundamental relationships about fidelity and transfer. The only way to evaluate their effectiveness is to try them.

AWAVS AS A CRITERION DEVICE

Although implementation is still a future consideration, planning might begin at this time regarding the use of AWAVS as a criterion device for transfer of training research. This means that a particular configuration of AWAVS, rather than an actual aircraft, would be used to evaluate transfer in pilot training studies. This approach is differentiated from that found in the "quasi-transfer" studies proposed earlier by the addition of an empirical data collection effort to effectively equate a simulator configuration to the aircraft. Only after an AWAVS configuration is so equated can experimental data with the simulator substituted for the aircraft be interpreted with confidence. One method of achieving this equivalence has been proposed by Matheny (1974). Some effort now might be devoted to a study to discover if such programs have ever been implemented (and if so, their current status), and whether they might be improved upon, particularly in regard to simplification and economy.

SECTION VII

ECONOMICAL DATA COLLECTION PLANS FOR TRANSFER OF TRAINING STUDIES FOR THE AWAVS PROGRAM

If the multifactor approach is to be applied to transfer of training research (as opposed to performance research) then it is necessary to find even more economical data collection plans that are suitable for this class of problem. The cost of data collection is intensified in a transfer of training study over that found in a performance study because each experimental condition is first associated with an extensive training period in the simulator and later tested in flight in the aircraft. Some ways of reducing this burden are suggested here. It should be noted, however, that these ideas are still in a conceptual stage, requiring empirical experience to test them and turn them into working plans, or to ultimately discard them.

Two basic approaches are proposed for economically discovering simulator configurations on which transfer effectiveness should be high. These are:

a. One in which a complete and thorough multifactor study of simulator factors is conducted using pilots skilled enough to fly the simulator without extensive training. This would be followed by a second, smaller and more limited transfer of training study using a second group of pilots with varying degrees of experience on the particular task, who will be given simulator training before performing in the aircraft. The performance measures from the first study would be related mathematically to the transfer occurring in the second study. The intent is to find an equation that will enable us to predict and safely extrapolate from the data least expensive to collect.

b. The other in which a transfer of training study is conducted (without a preliminary performance experiment) using economical multifactor data collection plans for the simulator training phase with equal or fewer conditions tested later in the flight phase. Economy is effected through the use of sequential data collection strategies and the reduction of in-aircraft tests.

In both approaches, the data collection effort is reduced (and economy is effected) as the costs in time, money, and difficulty of each phase -- performance, training, or flight -- increases. A fundamental assumption in the proposed approaches is that a poor model of a relatively complete multifactor study will give more accurate and useful prediction data than a better model of a severely limited part of the overall space.

A number of specific plans associated with each approach can be conceptualized as shown in Table 8. They obviously vary in cost and probably effectiveness. Which one would be used depends upon the circumstances at the time, i.e., the available resources (men and equipment), time, and above all the dedication of those involved to the research effort.

PERFORMANCE TO TRANSFER APPROACH (I)

These approaches all use the results of a complete multifactor performance study to select or otherwise minimize the number of conditions that need be included in a transfer of training study.

Selected Configurations (Plan 1-A)

A complete multifactor performance experiment would be performed first in the simulator. Pilots would be used who were sufficiently skilled to minimize an extensive training period in order to fly the simulator. They would, however, fall into at least two groups with high and low experience in making carrier landings (or whatever the experimental task may be). Two or more levels of task difficulty would also be included. Multifactor systematic screening designs in the paradigm proposed by Simon (1977) would be used for this study to make the data collection as economical as possible. Multiple performance measures (i.e., dependent variables) relevant to the task which could also be measured in the aircraft would be used. Additional measures might also be taken.

Next, a classic transfer of training experiment would be performed independently of the performance study. New pilots would be selected, with minimum carrier landing experience, but with one group being high skill pilots and another being low skill pilots. They would all be trained first in the simulator and later tested in the aircraft in flight.

The particular configurations to be used in the transfer experiment would be based on a study of the results of the performance experiment. For Plan 1-A no other use of the performance data (as it relates to the transfer study) is planned. The purpose of this approach is to limit the number of configurations to be used in the transfer of training study to only the most interesting. The exact number depends on the resources, the information desired, and any formal requirements of the experimental design.

Criteria for selecting particular configurations might include:

a. Performance level achieved. Configurations on which high, low, and medium performance levels were achieved might be selected to see to what extent transfer effectiveness correlates with performance effectiveness.

TABLE 5. APPROACHES TO ECONOMICAL TRANSFER OF TRAINING RESEARCH

I. RELATING PERFORMANCE TO TRANSFER

	Performance Studies	Transfer of Training Studies (Simulator Tng. - - → Aircraft Flight)
PILOT TYPES	High Skill Pilots a) Minimum carrier landing experience b) Maximum carrier landing experience	Minimum Carrier Landing Experience Pilots a) High skill pilots b) Low skill pilots
APPROACH	Collect the data required to develop a full model multifactor simulation performance map.	PLAN 1-A. Select a few simulator configurations to investigate on the basis of an examination of the performance map. Include additional configurations of practical and scientific interest related to simulator design problems. With these do a classic transfer study. PLAN 1-B. Systematically develop a transfer response surface using a fractional factorial design of low resolution. Use it to provide criteria for writing a transfer-prediction equation from performance data. Validate.

(continued)

TABLE 8. APPROACHES TO ECONOMICAL TRANSFER OF TRAINING RESEARCH (cont.)

II. PERFORM LIMITED DIRECT TRANSFER OF TRAINING STUDIES

	Transfer of Training Studies
PILOT TYPES	The low skill pilots with no carrier landing experience (or whatever population the transfer data is to be generalized to): all receive simulator training plus flight time.
APPROACH	PLAN II-A. Use 'new paradigm' for economical data collection to empirically map a transfer surface, building sequentially each data point involving both training and flight time, until an adequate model is represented.
	PLAN II-B. Use evolutionary operational approach to search for configuration yielding maximum transfer.
	PLAN II-C. Develop full model, multifactor training map, but limit continuations to flight to a minimum fractional factorial design. Use the flight data to provide criteria for writing a transfer-prediction equation from training data. Validate.

b. Engineering cost advantages. How much do simulator configurations on which performance levels are practically the same but which differ considerably in production costs differ in transfer effectiveness? Is an increase in transfer actually cost effective from an engineering point of view?

c. Engineering state-of-the-art advantages. Some configurations produce reasonably adequate simulation and acceptable in-simulator performance levels without straining the state-of-the-art. Other configurations may require additional engineering development to advance the state-of-the-art but may be less reliable and more costly to operate or maintain. How do they differ in regard to transfer effectiveness?

d. Correspondence with reality. To what extent does "fidelity" of simulation affect transfer effectiveness? If we select configurations that approximate reality well and not well, is transfer effectiveness markedly different in the two cases?

e. "Scientific" knowledge. The investigator might include any configurations that might increase his understanding of the transfer process, particularly as to how it relates to the performance effectiveness.

It is not possible to list all the detailed questions that might be investigated. They will have to be determined by the pattern of the performance response surface, the imagination and curiosity of the investigator, as well as his knowledge of the problem. Furthermore, such decisions will be limited by the time and money available for the follow-up transfer study.

The five criteria listed above are probably not completely orthogonal. For example, the most realistic configuration might be the most costly, the most complex, and the most unreliable. Still, they are representative of things an investigator may wish to explore for the transfer problem.

The main advantage of this plan is that, since it is not factorial in design, it does not place restrictions on the number or composition of the configurations (conditions) that will be examined. As many configurations as the investigator wishes would be examined in a simple one-way RMWR design; each configuration is treated as a qualitatively different condition. The disadvantages of this plan are 1) it requires redundant information to be collected; each condition must be repeated a sufficient number of times to provide some reliability to the means, and 2) the manner in which configurations are selected increases the chance that important configurations will be overlooked and important relationships missed. This approach, at best, is a make-shift one, and certainly the one least likely to be effective. It is expedient, but where long range planning is possible, some other approach should be employed.

Performance-Transfer Prediction Approach (Plan 1-B)

The purpose of this approach is to develop an equation that would predict transfer effectiveness from simulator performance measurements. If a valid prediction equation could be established, the performance data, which is less expensive to collect, would be used to estimate the transfer effectiveness of configurations not actually studied in the experiment and possibly of other simulator configurations involving similar tasks.

With a group of skilled pilots, a multifactor simulator performance study would be performed. Since excessive simulator training would not be required, this phase of the plan should be as complete as possible. This performance experiment is identical in process and result with that obtained in Plan 1-A.

The transfer phase of the plan would differ from Plan 1-A. Training and flight tests would be conducted by a different pilot on each configuration, but the configurations would be selected in a systematic manner to take advantage of the economy offered by internal replication rather than redundant replication of the same conditions. The intent would be to employ a minimum fractional factorial plan to create a transfer map over the same experimental space that had been covered by the performance map. However, the transfer map would be represented by a lower-order equation, and might only roughly approximate the true transfer response surface.

Transfer data would be collected in a series of small blocks, i.e., different small fractions of a total factorial. As each block is collected, the sum total of data up to that point would be correlated with the complete data from the performance maps to see how strong a relationship could be found. Presumably as the transfer response surface is more completely approximated, the more likely the relationship between it and the performance response surface can be used for prediction purposes. However, the intent is to stop before too much transfer data has been collected when additional improvement seems unjustified for prediction purposes. The assumption is made that even a poor approximation of a rather complete multifactor transfer surface will ultimately enable a better prediction -- operationally -- from performance data than were a limited transfer surface approximated.

At least two methods of relating the performance and transfer data might be tried: 1) to correlate only the responses from corresponding configurations in both sets of data; 2) to first use the collected transfer data to estimate transfer values at configurations at which no empirical data had been collected but which correspond to configurations used in the

performance study; then correlate estimated and empirical transfer data with performance data. Other reasonable variations on these techniques could be tried. The proportion of variance by which the two sources of data overlap serves as an indication of the strength of their relationship.

Several conditions might operate to make the relationship between the two sources low. One, the model of the equation from the transfer data may not -- because of the small amount of data allotted to that segment of the investigation -- be complex enough order to make accurate estimates. The sequential approach, however, would allow the model to be built a block at a time until it is optimized if the time and money available permits it. Two, there may be enough data, but in the wrong metric scale; data transformations would be required. Three, other factors than simulator performance may affect the level of performance in the aircraft and the transfer effectiveness measure. This means that once no further increment in the relationship can be achieved by enhancing the model of the transfer data, the investigator will want to look for other factors such as simulator fidelity and task difficulty that might account for unexplained sources of variance. An important part of this study would be the validation of the derived question. The transfer effectiveness of other configurations would be predicted and the prediction checked empirically.

Another factor that might account for the low relationship, if one is found, is the difference in the pilot populations that were used to get the performance and the transfer data. Ordinarily more skillful pilots may be used in the simulator performance study when minimum training is involved than in the transfer study where extensive training may be needed. Problems of interpretation might arise if configuration-by-pilot skill/experience interactions were to occur but could not be isolated. Therefore, unless all pilot combinations are to be included, pilots from the same populations should be used for the performance and transfer phases in this approach.

Since we have had no experience calculating these relationships, we must be prepared for them to be low. While it seems reasonable to expect some kind of relationship to exist, even with other intervening, covariant factors, it may not be sufficient for prediction purposes. If it turns out that no relationship can be established, that itself would be an important finding.

LIMITED DIRECT TRANSFER APPROACH (II)

In this approach, no initial performance study would be performed. Instead, we would start immediately with a multifactor transfer of training experiment using the strategy for economical data collection described earlier for the construction of a performance map.

Complete Transfer Surface (Plan II-A)

For a given pilot population and task, each pilot would be trained on a particular simulator configuration, after which his performance in the air would be tested. The transfer effectiveness of each pilot/configuration combination would be calculated separately. The simulator configurations, representing experimental conditions, would be selected and used according to the "new paradigm" described for economical multifactor research by Simon (1977b). To keep the study as inexpensive as possible, the principles of sequential data collection would be employed, starting with minimum-order designs, and progressing until the model adequately fits the responses. However, the intent of this plan is to create a complete transfer surface.

Initially, simulator configurations would be selected to provide a Resolution III design. Theoretically, we can study the effects of N simulator factors with N + 1 pilot/configurations if there are no interaction effects among factors. Since two factor interactions are common in behavioral research, the investigator will probably continue the data collection on (N + 1) new pilot/configurations in order to isolate main from two factor interaction effects. Of course, inspection of the first block of data may negate or modify the second step. After an inspection of the new data (combined with that from the first block), the investigator may wish to add other configurations to determine a second order response surface (if the factors are quantitative and continuous and if that accurate a representation is justified). The investigator always has the option of continuing or stopping.

The advantages of this approach are that it is direct, relatively uncomplicated, and the most economical way of collecting data for the amount of information indicated. Since each data point is collected independently of the others -- a different pilot/configuration being used on each -- scheduling and other logistic problems are simplified. The immediate information obtained is a measure of transfer effectiveness and the response surface is a transfer surface.

The disadvantage of this approach is that although the paradigm is the least costly data collection plan for the amount of information obtained, being a transfer of training study, it is still expensive. This approach does not offer the opportunity to develop equations that might permit predictions to be made from prior performance studies, which can usually be done far more economically than a transfer study, but possibly not as accurately.

Search-For-Optimum Transfer (II-B)

If many of the critical simulator factors were quantitative and it were possible to examine configurations at continuous points between the extreme ranges of interest, then a search strategy employed in industry to optimize production yields (EVOP) might be used to search for the most transfer-effective configuration. While many factors may not be quantitative or continuous, there are usually sub-groups that can meet this qualification which might be investigated in a separate study once a more gross, overall transfer pattern has been developed.

The basic experimental design begins much like a screening design, being a Resolution III 2^{k-p} design. The main difference is in the range that is covered by each parameter of the design. In screening designs, one tries to encompass the total effective operational range of the experimental variables immediately. The assumption is made that this can be estimated and that the relationships within those limits can ordinarily be approximated by a second degree polynomial. In search designs, the investigator starts by looking at only a small part of the total space of interest. He tries to guess where an optimum might be, but he does not attempt to cover the total range. Instead, he looks at a part of the total space and uses that data to estimate where to look next, each time approaching closer to where the optimum configuration for maximum transfer would lie. This continues until he locates it. The method would be used when the surface is too complex to be covered by a single design and the investigator has little idea of where the optimum might be.

A transfer of training study -- simulator training and aircraft test -- would be performed at the minimum number of conditions (i.e., simulator configurations) required to include all factors in a Resolution III fractional factorial design. Either the Box and Hunter or Plackett and Burman plans might be used, the latter in some cases requiring fewer data points. The space encompassed by the experimental points would be only a small part of the total space of operational interest. A different pilot would be tested on each condition. The results of this initial data collection effort, in the form of a first order polynomial, would be used to estimate the direction, away from the space covered by the original study, in which the configuration yielding the greatest amount of transfer is likely to be found. (This, of course, assumes that it is not within the space originally examined.) A second set of observations (Resolution III) would be made at new coordinates in that vicinity. This procedure would be repeated until the observations appear to surround the location of maximum transfer.

A disadvantage of this plan, when it can be used, is that it seeks a point of optimum transfer. Seldom in human factors work is a single point sufficient information, since design decisions must often be compromises among performance, costs,

and other practical considerations. At times, optimum results may take the form of a ridge of equal performance, in which case, some trade-offs could be made. Were the effort worth it, a response surface might be completed for the space around the optimum point. Additional data would have to be taken to fit the surface to the model correctly representing the complexity of the surface.

Reduced-Flight Predicted Transfer (II-C)

In this approach, economy is achieved by reducing the amount of flight data that would be required. This would be accomplished in one of two ways: 1) to predict transfer effectiveness of simulator configurations that were never flight-tested by using equations representing the response surfaces that were derived from transfer data (based on training and flight test) made on only a few configurations; 2) to predict transfer effectiveness from performance data collected during the training period after the relationship between training performance and flight performance has been established. These two approaches employ features that are similar to the Search Approach and to the Performance-Transfer Prediction Approach, respectively.

In both cases, complete transfer studies would be performed on the configurations making up a Resolution III design. If time and money limitations permit, a higher resolution design would be employed involving more experimental conditions. Training performance data would be obtained, followed by the flight test data. Transfer effectiveness values could be calculated for all of these configurations and a first order, linear polynomial could be written from the transfer data that could be used to predict transfer effectiveness for other configurations. How accurate this prediction would be depends on how well the equation approximates the response surface. If one must extrapolate beyond the boundaries of the original experiment, predictions could be quite inaccurate.

With that data from this limited study, however, the investigator would have a second means of estimating transfer effectiveness. He could take performance measures collected at different stages of the training phase and see how they correlate with performance in the aircraft (or transfer effectiveness). This correlation, as an equation, could also be used to predict transfer effectiveness for other configurations provided the training data were made available on those configurations.

Of course, these descriptions of both techniques are oversimplified. It is unlikely that high correlations will occur without additional work on the part of the investigator. Quite probably other parameters, e.g., fidelity, task difficulty, pilot skill/experience, would have to be introduced as multiple predictors to improve the estimates of transfer effectiveness.

Perhaps the two measures combined into a single equation might provide a more accurate prediction. It may be that the prediction is only suitable for ranking a set of configurations but not for measuring the actual amount of transfer. These are all experimental questions that can only be answered empirically.

SECTION VIII

SOME UNFINISHED BUSINESS - MEASUREMENTS AND CRITERIA

Certain questions associated with performance and transfer measures remain unanswered although the answers to each affect, to some extent, the usefulness of the proposed methodologies as well as the very effectiveness of the AWAVS human performance research program.

First, there is the question of what performance measures will be taken on both the physical system and the pilot system? The usefulness of the experimental results depends on how relevant the measurements made in the experiment are to the operational task. More numbers, taken because they are more expedient or convenient, do not guarantee that the results of the study will be useful or even correct insofar as the operational situation is concerned. Will performance data be available to the investigator during a run, within moments following the run, by the time a second pilot is to be run, or when? Will there be the capability of performing summary analyses on the raw data? How quickly might that be available? Advanced experimental methods are economical because of their sequential nature. That means that they rely on a process whereby a small block of data is collected and examined (analyzed) to determine if and what subsequent steps are needed. If this process is delayed beyond the time it takes to set up for the next trial, the data collection period is not only drawn out inefficiently but the effects of the delay on the pilot could conceivably distort his performance.

Another problem related to measurement in a transfer of training study involves the criterion of training employed. Will the interpretation of the results differ if we use time-to-criterion, or if we use equal number of training trials, or if the criteria we employ (as we should) are multiple response measures? Associated with these questions are others, such as: how does the use of different criteria affect the reliability of the results, the logistic problems of running the experiment, and so forth?

A third problem related to measurement has to do with the preferred form of measurement to be employed in the analysis. While we are ultimately interested in transfer effectiveness, data expressed in those terms are in fact particular transformations of performance scores. It is necessary to discover whether predictions might be more easily and accurately made if more basic performance measures were employed, leaving particular transformations up to the users of the data. For example, we may find that performance in the aircraft can be predicted from performance in the simulator more readily than transfer measurements in the aircraft. Then again, we may not.

Measurement problems are fundamental to any research conducted on transfer of training and to ignore them or assume that previous research has resolved these questions can only increase the risk that our experimental efforts will fail.

REFERENCES

Addleman, S., Techniques for constructing fractional replicate plans, J. Amer. Statis. Assoc., 1963, 58, 45-71.

Caro, P. W., Aircraft simulators and pilot training, Human Factors, 1973, 15, 502-509.

Daniel, C., Sequences of fractional replicates in the 2^{p-q} series, J. Amer. Statis. Assoc., 1962, 57, 403-429.

Daniel, C., Application of statistics to industrial experimentation, N. Y.: Wiley, 1976.

Kron, G. L., Fidelity of simulation, Binghamton, N. Y.: Link Division/Singer Company, September 1970.

Matheny, W. G. Training research programs and plans (ASUPT), Williams AFB, AZ: LSI-TR-74-2, November 1974.

NAS-NRC Vision Committee, Visual cues in flight simulation, (Draft copy), 16 September 1976.

Naval Training Equipment Center, Aviation wide angle visual system (AWAVS) utilization plan, Orlando, Fla.: NTEC, May 1978.

Roscoe, S. N., Effective and economical simulation in the design and use of aero systems, Urbana-Champaign: University of Illinois ARL-75-8/AFOSR-75-3, April 1975.

Simon, C. W. Reducing irrelevant variance through the use of blocked experimental designs. Culver City, CA: Hughes Aircraft Co., Tech. Rep. No. AFOSR-70-5, November 1970a, 65pp. (AD 776-041)

Simon, C. W. The use of central-composite designs in human factors engineering experiments. Culver City, CA: Hughes Aircraft Co., Tech. Rep. No. AFOSR-70-6, December 1970b, 52 pp. (AD 748-277)

Simon, C. W. Considerations for the proper design and interpretation of human factors engineering experiments. Culver City, CA: Hughes Aircraft Co., Tech. Rep. No. P73-325, December 1971, 135 pp.

Simon, C. W. Experiment simulation. Culver City, CA: Hughes Aircraft Co., Tech. Rep. No. ARL-72-7/AFOSR-72-3, April 1972, 48 pp. (AD 754-215)

Simon, C. W. Economical multifactor designs for human factors engineering experiments. Culver City, CA: Hughes Aircraft Co., Tech. Rep. No. P73-326A, June 1973, 171 pp. (AD A035-108)

Simon, C. W. Methods for handling sequence effects in human factors engineering experiments. Culver City, CA: Hughes Aircraft Co., Tech. Rep. No. P74-451A, December 1974, 197 pp. (AD A035-109)

Simon, C. W. Methods for improving information from "undesigned" human factors experiments. Culver City, CA: Hughes Aircraft Co., Tech. Rep. No. P75-287, July 1975a, 82 pp. (AD A018-455)

Simon, C. W. Evaluation of basic and applied research. Pragmatic criteria. Paper presented at 83rd Annual Convention, American Psychological Association, Chicago, IL, 31 August 1975b, 20 pp.

Simon, C. W. Response surface methodology revisited: a commentary on research strategy. Westlake Village, CA: Canyon Research Group, Inc., Tech. Rep. No. CWS-01-76, July 1976a, 60 pp. (AD A043-242)

Simon, C. W. Analysis of human factors engineering experiments: characteristics, results and applications. Westlake Village, CA: Canyon Research Group, Inc., Tech. Rep. No. CWS-02-76, August 1976b, 104 pp. (AD A038-184)

Simon, C. W. Design, analysis and interpretation of screening designs for human factors engineering research. Westlake Village, CA: Canyon Research Group, Inc., Tech. Rep. No. CWS-03-77A, September 1977a, 220 pp. (AD A056-985)

Simon, C. W. New research paradigm for applied experimental psychology: a system approach. Westlake Village, CA: Canyon Research Group, Inc., Tech. Rep. No. CWS-04-77A, October 1977b, 123 pp. (AD A056-984)

Williams, A. C., Jr. and Adelson, M. Some considerations in deciding about the complexity of flight simulation. San Antonio, Lackland AFB. Research Bulletin AFPTRC-TR-54-106, December 1954.

Yates, F. The design and analysis of factorial experiments. Harpenden, England: Imperial Bureau of Soil Science, Tech. Commun. No. 35, 1937.

GLOSSARY

ALIAS — Screening and other fractional factorial designs (see below) do not isolate all main and interaction effects from one another. A comparison which intends to isolate one effect may therefore also include estimates of others. When two or more effects are 100 percent confounded in this way, the effects are said to be aliased. The estimated effect is actually the sum effect of the aliases. (See also, FRACTIONAL FACTORIAL DESIGN; CONFOUNDING).

CHANGE-OVER DESIGNS — (Sometimes referred to as carry-over, cross-over, or residual designs). These experimental designs are used when a subject is tested sequentially over a number of experimental conditions. These designs are constructed so as to isolate the direct effect of a treatment from any residual effect that may have been "carried over" from the previous treatment. Change over designs are distinguished from serially balanced sequence designs in that the necessary balance required to isolate direct and residual effects is distributed among a number of subjects in the change-over design but is complete within a single subject for the serially balanced sequence design. (See also, SERIALLY BALANCED SEQUENCE DESIGN).

CONFOUNDING — When estimates of the effects of two or more sources of performance variance cannot be completely isolated, either intentionally or through faulty experimental design, the effects are said to be confounded. Confounding may range from some minimal percent up to 100 percent. (See also ALIAS).

FACTOR LEVEL CHANGE NUMBER When experimental conditions are run sequentially, the level or setting of each factor must be changed from time to time. In screening designs, the "change number" indicates the total number of times a particular factor must be switched between its high and low levels. It is important in the design of an experiment when making the change is difficult or otherwise costly. (See also SCREENING DESIGN).

FRACTIONAL FACTORIAL DESIGN This experimental design is composed of some fractional subset of the total number of experimental conditions in the complete factorial. It is employed when certain effects (generally higher-order interactions) are expected to be negligible or nonexistent. Subsets of experimental conditions for the fraction are selected in a way that allows the comparison for the negligible effects to be used to measure the effects of additional factors aliased with them. Fractional factorials of two levels are commonly designated in the form 2^{k-p}. For example, a 2^{8-4} fractional factorial would be a 2^{4} or a 1/16 fraction of a complete 2^{8} factorial. That is, a particular subset of 16 conditions out of a total of 256 would be used to study eight factors at two levels each. A "saturated" fractional factorial design is one in which there are n observations for n-1 main effects.

HALF-NORMAL PLOTS This graphic technique is used to identify visually the critical effects of 2^{k} factorial or 2^{k-p} fractional factorial experiments that have been plotted in order of absolute magnitude on half-normal plotting paper. (See also FRACTIONAL FACTORIAL DESIGN).

HOLISTIC	A philosophic point of view in the conduct of behavioral experiments that emphasizes the importance of accounting for as many critical variables as possible, whether equipment, environment, subject, or temporal, controlled or uncontrolled. Implementing such a philosophy requires the application of principles of economical multifactor designs. (See also REDUCTIONISTIC).
MULTIFACTOR EXPERIMENT	As used in this report, a multifactor experiment is one which attempts to satisfy the holistic philosophy. Thus, a three or even five factor experiment (at the beginning of a research program), while involving multiple factors, would not necessarily be a multifactor experiment as the term is used here. Compromises with non-experimental conditions surrounding an experiment may make it impossible to include all potentially critical factors, but the initial emphasis will be on trying to do so. (See also HOLISTIC).
ORTHOGONALITY	That property of an experimental design which insures that the different effects shall be capable of [illegible] and separate estimation without any confounding. The sums of squares of all effects will be independent and additive. (See also CONFOUNDING).
PERFORMANCE EXPERIMENT	As the term is used in this report, a performance experiment is one that measures operator/system performance under one set of conditions, presumably unaffiliated by any other prior conditions. Measuring pilot performance in a simulator with different configurations could be an example of this type, as opposed to another type referred to as a "transfer" experiment. (See also TRANSFER EXPERIMENT; QUASI-TRANSFER EXPERIMENT).

PRINCIPLE OF PROPORTIONAL PROPERTIES	A necessary and sufficient condition that the main effects of two factors be uncorrelated is that the levels of one factor occur with each of the levels of the other factor with proportional (not necessarily equal) frequency.
QUASI-TRANSFER EXPERIMENT	This is a transfer experiment in which performance is never measured under realistic, i.e., non-simulation, conditions. For pilot training, this means that the experiment would include no post-training period in which performance was measured in the aircraft. Instead an alternate simulation configuration would be employed to represent flight conditions. (See also TRANSFER EXPERIMENT) (PERFORMANCE EXPERIMENT).
MONOTHETIC	A philosophic point of view in the conduct of behavioral experiments that advocates reducing the variables in an experiment to the smallest number possible. In its extreme form the resulting experiment is one in which a single factor is varied and all other sources of variance are held constant. This philosophy is in direct opposition to the holistic philosophy. (See also HOLISTIC).
RESOLUTION	A design of "resolution" R is one in which no p-factor effect is confounded with any other effect containing fewer than R-p factors. The resolution of a design is noted by the appropriate Roman numeral as a subscript in the fractional factorial designation, e.g., $\frac{2^{k-p}}{2^{k}}$ design. A design of Resolution III does not confound main effects with one another, but does confound them with two-factor interactions. A design of Resolution IV

RESOLUTION (Continued)

isolates main effects from one another and from two-factor interactions, but the two-factor interactions are aliased in strings. A design of Resolution V isolates all main effects and all two-factor interactions from one another. In all screening designs, main effects and two-factor interactions are confounded with higher order effects. (See also, FRACTIONAL FACTORIAL DESIGN; SCREENING DESIGN).

RESPONSE SURFACE METHODOLOGY

This refers to a particular strategy introduced and promoted by G.E.P. Box and associates for conducting experiments to obtain an equation representing the response multifunction, or surface. It is not a design, per se, but the judicious use of principles of blocking, fractional factorials, and tests of model adequacy in a way that insures an accurate representation of performance within the experimental space at minimal data collection cost.

SCREENING DESIGN

As used in this report, it refers to a saturated or nearly saturated fractional factorial design capable of handling a large number of factors. These designs are all of the form, 2^{k-p}, generally of Resolution III or IV. The initial information is first evaluated before subsequent data are collected, the purpose being only to identify the critical factors within a larger candidate group. Additional data must be collected ordinarily to meet a second and separate purpose, defining the response surface. (See also, FRACTIONAL FACTORIAL DESIGN; RESOLUTION).

SERIALLY BALANCED SEQUENCE DESIGN	A modified change-over design for isolating direct and residual effects, in which the necessary balance occurs within the extended number of trials run by a single subject. This contrasts with the change-over design in which the balance is obtained among several subjects each tested on fewer trials. (See also CHANGE-OVER DESIGNS).
SINGLE FACTOR EXPERIMENT	This refers to the type of experiment proposed by the Reductionist. As used in this report, it need not be for one factor, but for any small number which is a seriously incomplete number of the potentially critical factors affecting the particular performance. (See also REDUCTIONISTIC).
TRANSFER EXPERIMENT	In contract with a performance experiment, as used here, this refers to experiments in which interest centers on the residual effects that practice on one set of conditions has on the performance of a second set which follows. (See also, PERFORMANCE EXPERIMENT; QUASI-TRANSFER EXPERIMENT).
TREND-ROBUST EFFECTS	Designs exist that isolate linear, quadratic, and/or cubic trend effects from experimental effects of interest. Examples of trend effects are subject learning, or equipment drift over time. A trend-robust effect is one which is not biased, or only minimally biased, by trends running through the data.